Analysis of the information system of the social services users (SIUSS) of the Málaga council

Rafael Hidalgo Calero

Abstract

PROPOSAL of a machine learning system that is able to predict the problems that affect to the people who go to social services in Málaga. For this, the data provided by the Observatorio Municipal para la Inclusión Social of the city council have been used. In the early stages of development has been implemented the process of extract, transform and load the data from the XML files provided to the Machine Learning system developed on the H_2O.ai platform. After that, we have proceeded to develop different Machine Learning algorithms trying to find those that are more useful to social workers. For this, numerous experiments have been carried out with different sets of data, parameters, models and objectives. Then, we have obtained a set of systems that are capable of making reliable predictions about the needs that new users of social services require. Finally, a web application has been developed to enable to introduce new input parameter and show the different predictions obtained.

Index Terms

Machine Learning, H_2O.ai, Social Services System, Málaga council, SparkSQL

I. Introduction

THIS project is about seeking to know the social and intervention needs that lead to avoid the social exclusion of people in the city of Málaga. This work is carried out within the framework of the agreement that the Faculty of Estudios Sociales y del Trabajo of the University of Málaga has with the City of Málaga through Observatorio Municipal para la Inclusión Social. Observatorio Municipal para la Inclusión Social is the operational tool of Área de Derechos Sociales designed to know the social reality of Málaga in general and the users of the public social services system in particular. This observatory consists of an interdisciplinary team of professionals in social work, psychology and computer science specializing in social and community research.

Goals:

- Provide the necessary information for decision making in social planning and intervention.
- Research on the causes and consequences of poverty and social exclusion.

After the first meetings with Observatorio Municipal para la Inclusión Social of the City of Málaga and, having known the methodology of work and tools that are used to study the different profiles of users of social services in the city, have been detected a priority development line that can be useful in their day-to-day, as well as provide new tools to improve the services offered by social workers to users.

It is proposed the development of a machine learning system that is able to predict the problems that affect a person who goes to social services. As well as, determine what are the best lines of action to take to help the quality of life of this person and their environment in the early stages of social services. In the same way, it can be a guide to help social workers to make better decisions. In short, it aims to provide a tool for the early detection of actions that social workers can perform to improve the quality of life of users of social services.

It is expected to establish profiles of users of social services, their needs and the type of intervention applied according to the urban, demographic and socioeconomic characteristics of the city.

Author: Rafael Hidalgo Calero, rafael.hidalgo.calero@gmail.com
Advisor 1: Luis Gómez Jacinto, Psicología Social, University of Málaga
Advisor 2: Antonio Jesús Nebro Urbaneja, Lenguajes y Ciencias de la Computación, University of Málaga
Thesis dissertation submitted: March 2018

II. State of the art

THERE is a great interest in knowing the social reality of the city of Málaga and the users of the social services of this city to improve knowledge of the most demanded needs and the problems that most affect the citizens. This allows providing better services to users and increasing the quality of life of the citizens of Málaga.

With this intention, the Observatorio Municipal para la Inclusión Social is responsible for carrying out studies based on the data stored in Sistema de Información de Usuarios de Servicios Sociales (SIUSS).

In this line, in 2014 was published Perfil de las personas usuarias de los SSAP de Málaga, 2013 [5], a study on the living conditions of the user population of primary care social services where the main results of the variables are presented analyzed for all the people of the families assisted during the year 2013.

This type of work and others carried out by the Observatorio Municipal para la Inclusión Social of Málaga are very important to know and understand the social reality of the city of Málaga and its evolution over time. However, these are descriptive studies that show what has happened or what is happening now, but that are not able to predict what will happen in the future, nor the needs that new users of social services will have.

Therefore, in this work we intend to perform a predictive analysis that is capable of anticipating the needs that users of social services of the city of Málaga require and thus offer a better service in the early stages of care.

III. Method

A. Data model

1) SIUSS database brief: Social service workers use Sistema de Información de Usuarios de los Servicios Sociales (SIUSS) as a tool for expedients management, a system developed by Salud, Servicios Sociales e Igualdad Spanish Ministry. It stores information about expedients, families user of social services, interventions, etc. The following are the main SIUSS data tables and on which the study has been carried out:

- TEXPFA: Family expedients. General information of each expedient
- TMIEMB: Members of the family unit
- TEQUEX: Equipment of the family unit related to the expedient
- TINTER: Interventions made for each expedient
- TUSUIN: User of the family unit receiving the intervention
- TVALIN: Specific evaluations of the intervention
- TDEMIN: Requests for intervention
- TRECID: Ideal resources to apply in the intervention
- TRECAP: Resources applied in the intervention
- TGESTINT: Management associated with the intervention

In appendix A, the detail of each table is specified

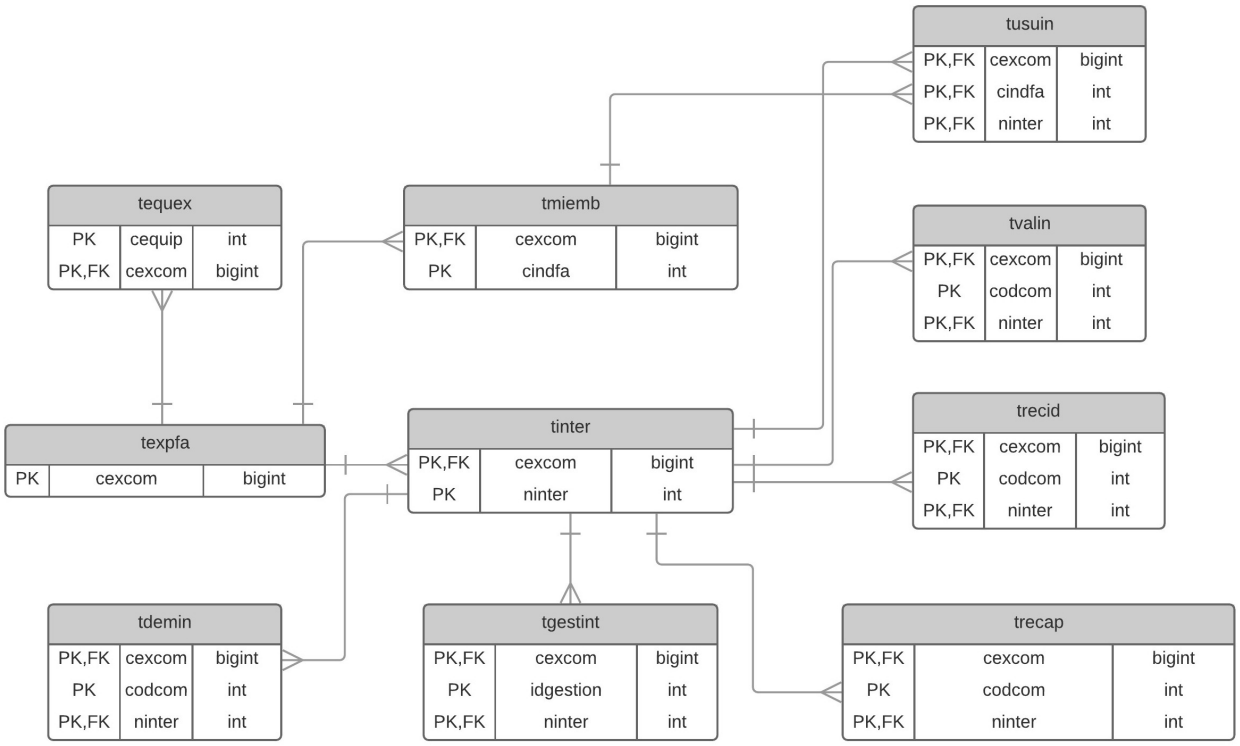


Fig. 1. SIUSS E/R diagram

When we are going to design our Machine Learning system, is important keep in mind that a family expedients is comprised for one or more interventions and each of one has multiple possibles values for evaluations, requests, ideal resources and applied resources. Also, an intervention is related with one or several member of family unit.

B. Selection of problems

Following detailed discussions, both the data model, described in the previous section, and possible needs for the social workers, a set of problems to address have been defined. From this definition, the processing of the data, building of the machine learning models and the application for operate with the system, have been built.

1) Predictions about evaluations: Tries to obtain the most probable evaluations for an intervention. As we said before, an intervention can have more than one evaluations, for that, the target is not get the better evaluation, just an ordered list of all possibles evaluation.

2) Prediction about applied resources: Tries to obtain the most probable applied resources for an intervention. Same as evaluations, an intervention can have more than one applied resource, for that, the result will be an ordered list of most probable resources to apply.

3) Prediction about ideal resources: Tries to obtain the most probable ideal resources for an intervention. Same as evaluations and applied resources, an intervention can have more than one ideal resource, for that, the result will be an ordered list of most probable ideal resources.

4) Prediction about if exist abuse: This problem is about getting if for an intervention, there is some kind of abuse in the family unit. There are just two possible values, exist or not abuse. For that, this is a binomial problem

5) Prediction about if applied resource matches the ideal resource: This problem is about getting if for an intervention, applied resources matches ideal resources. In this case, there are only two possible values, applied resources matches ideal resources or not. For that, this is a binomial problem.

C. Features selection

1) Removing of columns: As a preliminary phase to the import of data into the Machine Learning system, data filtering has been done, removing certain columns that are not going to add value in the learning process. The reasons why certain columns have been removed are listed below.

- Anonymized: There are anonymized columns, which contain confidential information and do not add value to the Machine Learning system (names, addresses, telephone numbers, etc.)
- Keys: Columns that act as primary or foreign keys, with auto-generated and non-informative values
- Descriptive fields: Columns that contain detailed information provided by the social worker. To be able to include this type of columns, a semantic analysis should be develop to categorize this information. This semantic analysis is out of the scope in this project but it would be a point to consider in future iterations.

Appendix B shows the list of removed columns and the reason.

2) Reducing of date values: In addition, all date values have been reduced to the year for simplicity and consider that they represent enough information for the system (e.g. 2017-01-08 10:45:32.000Z is transform to 2017).

3) Binarization of family equipment: For each family expedient, there are multiples values that indicate the home equipment, stored in TEQUEX table. For each possible value of family equipment, a new column has been created. If the value of these columns are 1, the family unit has this equipment, in other case, the value is 0.

In appendix C, the possible values of equipment is specified

4) Reduce to main member of family unit: In both tables, "TUSUIN" and "TMIEMB", there is a column that indicate if this member of the family unit is the main member ("CINDFA" = 1). To reduce the number of cases and simplify the dataset to be generated, only the main member of the family unit will be consider.

D. Auxiliary database

To make easier the system building, a auxiliary MySQL database has been created. Also a Spring Boot application has been implemented to interact with this database. The follows tables have been created:

- TFIELD: Stores all columns for all SIUSS tables used, also, stores generated columns (binarization of family equipment, if exist or not abuse and if applied resource matches ideal resource)
  - id: Identifier of the field
  - description: Description of the field
  - field: Name of the field
  - field_translate: Transformation to apply for this field ("reduce_to_year" to date fields)
  - field_type: Type of the field. The possible values are: description, code, numeric and label
  - owner: Owner table of the field
  - visible: Indicates if this field is included in the system
  - name: Final name of the field in the generated datasets. owner + "_" + field
  - label: Descriptive name of the field, used in the web application
- TYPOLOGY: Stores all possible values of some fields of code type. This is used to show a combo box in the web application
  - id: Identifier of row
  - field: Name of the owner field
  - cod: Code of the resource
  - description: Description of the typology
  - depend_on: Identifier of the parent typology
- TRECURSOS: Stores all possible values for resources, it is valid for applied and ideal resources
  - cod: Code of the resource
  - descripcion: Description of the resource
- TVALORACIONES: Stores all possible values for evaluations
  - cod: Code of the evaluation
  - descripcion: Description of the evaluation

Furthermore, these are REST services created to obtain information from auxiliary database:

TABLE I
SIUSS GET FIELD SERVICE

Endpoint	**/siuss/data/field**
Description	Get information about all fields. Field name, description, type of data, is visible, transform to apply, etc...
Type	GET
Input	-
Output	List of oscuroweb.bigdata.dto.FieldDTO.

TABLE II
SIUSS GET VISIBLE FIELD SERVICE

Endpoint	**/siuss/data/field/visible**
Description	Get information about all visible fields
Type	GET
Input	-
Output	List of oscuroweb.bigdata.dto.FieldDTO.

TABLE III
SIUSS TYPOLOGY

Endpoint	**/siuss/data/tipology**
Description	Get information about all typologies
Type	GET
Input	-
Output	List of oscuroweb.bigdata.dto.TypologyDTO.

TABLE IV
SIUSS TYPOLOGY BY FIELD

Endpoint	**/siuss/data/tipology/{field}**
Description	Get information about all typologies by field
Type	GET
Input	field: String
Output	List of oscuroweb.bigdata.dto.TypologyDTO.

TABLE V
SIUSS TYPOLOGY BY PARENT

Endpoint	**/siuss/data/tipology/**
Description	Get information about all typologies by parent value
Type	POST
Input	parent: oscuroweb.bigata.dto.TypologyDTO
Output	List of oscuroweb.bigdata.dto.TypologyDTO

TABLE VI
SIUSS EVALUATION BY CODE

Endpoint	**/siuss/data/valoracion/{code}**
Description	Get information about a evaluation by code
Type	GET
Input	code: String
Output	oscuroweb.bigdata.dto.ValoracionDTO.

TABLE VII
SIUSS RESOURCE BY CODE

Endpoint	**/siuss/data/recurso/{code}**
Description	Get information about a resource by code
Type	GET
Input	code: String
Output	oscuroweb.bigdata.dto.RecursoDTO.

E. Extract, Transform and Load

Observatorio Municipal para la Inclusión Social of Málaga has grant a set of SIUSS data large enought to the development of the Machine Learning system. This set of data is represented in 11 XML files, where each of them contains the information related to a district of the city of Málaga.

This is the XML format:

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<Salida5>
  <TABLE1>
      <FIELD1>value</FIELD1>

      <FIELDN>value</FIELDN>
  </TABLE1>

  <TABLEJ>
      <FIELD1>value</FIELD1>

      <FIELDM>value</FIELDM>
  </TABLEJ>
</Salida5>
```

Therefore, the first step for the construction of the Machine Learning system will be to extract the data from the XML files and transform them into a format that can be loaded on the H_2O.ai platform.

To make this ETL process, an Apache SparkSQL service has been developed. This service has the following steps:

- Get the list of fields: Invoke service siuss/data/field mentioned in previous section

```
fields = Arrays.asList(restTemplate.getForEntity("http://localhost:8081/siuss/data/field",
    FieldDTO[].class).getBody());
```

- Obtain XML files and load in org.apache.spark.sql.Dataset: For each SIUSS table, load data from XML and store in a Dataset instance

```
Dataset<Row> dataset = spark.read()
                .format("com.databricks.spark.xml")
                .option("ignoreLeadingWhiteSpace", true)
                .option("rowTag", rowTag)
                .load(csvFile)
                .toDF();
```

- Rename datasets columns name: For each column, rename it with TABLE_NAME + "_" + COL_NAME

```
String[] columns = dataset.columns();

Dataset<Row> renamedDataset = dataset;
for (String col : columns) {
  renamedDataset = renamedDataset.withColumnRenamed(col, colName(table, col));
}
```

- Make a binarization of family equipment data and group by expedient code

```
Dataset<Row> tequexBinDataset = tequexDataset
            .withColumn("TEQUEX_CEQUIP1", tequexDataset.col("TEQUEX_CEQUIP").equalTo("1").
                cast("integer"))
            .withColumn("TEQUEX_CEQUIP2", tequexDataset.col("TEQUEX_CEQUIP").equalTo("2").
                cast("integer"))
            .withColumn("TEQUEX_CEQUIP3", tequexDataset.col("TEQUEX_CEQUIP").equalTo("3").
                cast("integer"))
            .withColumn("TEQUEX_CEQUIP4", tequexDataset.col("TEQUEX_CEQUIP").equalTo("4").
                cast("integer"))
            .withColumn("TEQUEX_CEQUIP5", tequexDataset.col("TEQUEX_CEQUIP").equalTo("5").
                cast("integer"))
            .withColumn("TEQUEX_CEQUIP6", tequexDataset.col("TEQUEX_CEQUIP").equalTo("6").
                cast("integer"))
            .withColumn("TEQUEX_CEQUIP7", tequexDataset.col("TEQUEX_CEQUIP").equalTo("7").
                cast("integer"))
```

```
            .withColumn("TEQUEX_CEQUIP8", tequexDataset.col("TEQUEX_CEQUIP").equalTo("8").
                cast("integer"))
            .withColumn("TEQUEX_CEQUIP9", tequexDataset.col("TEQUEX_CEQUIP").equalTo("9").
                cast("integer"))
            .withColumn("TEQUEX_CEQUIP10", tequexDataset.col("TEQUEX_CEQUIP").equalTo("10")
                .cast("integer"))
            .withColumn("TEQUEX_CEQUIP11", tequexDataset.col("TEQUEX_CEQUIP").equalTo("11")
                .cast("integer"))
            .withColumn("TEQUEX_CEQUIP12", tequexDataset.col("TEQUEX_CEQUIP").equalTo("12")
                .cast("integer"))
            .withColumn("TEQUEX_CEQUIP13", tequexDataset.col("TEQUEX_CEQUIP").equalTo("13")
                .cast("integer"))
            .withColumn("TEQUEX_CEQUIP14", tequexDataset.col("TEQUEX_CEQUIP").equalTo("14")
                .cast("integer"))
            .withColumn("TEQUEX_CEQUIP15", tequexDataset.col("TEQUEX_CEQUIP").equalTo("15")
                .cast("integer"))
            .withColumn("TEQUEX_CEQUIP16", tequexDataset.col("TEQUEX_CEQUIP").equalTo("16")
                .cast("integer"))
            .withColumn("TEQUEX_CEQUIP17", tequexDataset.col("TEQUEX_CEQUIP").equalTo("17")
                .cast("integer"))
            .select("TEQUEX_CEXCOM", "TEQUEX_CEQUIP1", "TEQUEX_CEQUIP2", "TEQUEX_CEQUIP3"
                , "TEQUEX_CEQUIP4", "TEQUEX_CEQUIP5", "TEQUEX_CEQUIP6"
                , "TEQUEX_CEQUIP7", "TEQUEX_CEQUIP8", "TEQUEX_CEQUIP9"
                , "TEQUEX_CEQUIP10", "TEQUEX_CEQUIP11", "TEQUEX_CEQUIP12"
                , "TEQUEX_CEQUIP13", "TEQUEX_CEQUIP14", "TEQUEX_CEQUIP15"
                , "TEQUEX_CEQUIP16", "TEQUEX_CEQUIP17");


        tequexBinDataset = tequexBinDataset.groupBy("TEQUEX_CEXCOM")
            .sum("TEQUEX_CEQUIP1", "TEQUEX_CEQUIP2", "TEQUEX_CEQUIP3"
                , "TEQUEX_CEQUIP4", "TEQUEX_CEQUIP5", "TEQUEX_CEQUIP6"
                , "TEQUEX_CEQUIP7", "TEQUEX_CEQUIP8", "TEQUEX_CEQUIP9"
                , "TEQUEX_CEQUIP10", "TEQUEX_CEQUIP11", "TEQUEX_CEQUIP12"
                , "TEQUEX_CEQUIP13", "TEQUEX_CEQUIP14", "TEQUEX_CEQUIP15"
                , "TEQUEX_CEQUIP16", "TEQUEX_CEQUIP17")
            .withColumnRenamed("sum(TEQUEX_CEQUIP1)", "TEQUEX_CEQUIP1")
            .withColumnRenamed("sum(TEQUEX_CEQUIP2)", "TEQUEX_CEQUIP2")
            .withColumnRenamed("sum(TEQUEX_CEQUIP3)", "TEQUEX_CEQUIP3")
            .withColumnRenamed("sum(TEQUEX_CEQUIP4)", "TEQUEX_CEQUIP4")
            .withColumnRenamed("sum(TEQUEX_CEQUIP5)", "TEQUEX_CEQUIP5")
            .withColumnRenamed("sum(TEQUEX_CEQUIP6)", "TEQUEX_CEQUIP6")
            .withColumnRenamed("sum(TEQUEX_CEQUIP7)", "TEQUEX_CEQUIP7")
            .withColumnRenamed("sum(TEQUEX_CEQUIP8)", "TEQUEX_CEQUIP8")
            .withColumnRenamed("sum(TEQUEX_CEQUIP9)", "TEQUEX_CEQUIP9")
            .withColumnRenamed("sum(TEQUEX_CEQUIP10)", "TEQUEX_CEQUIP10")
            .withColumnRenamed("sum(TEQUEX_CEQUIP11)", "TEQUEX_CEQUIP11")
            .withColumnRenamed("sum(TEQUEX_CEQUIP12)", "TEQUEX_CEQUIP12")
            .withColumnRenamed("sum(TEQUEX_CEQUIP13)", "TEQUEX_CEQUIP13")
            .withColumnRenamed("sum(TEQUEX_CEQUIP14)", "TEQUEX_CEQUIP14")
            .withColumnRenamed("sum(TEQUEX_CEQUIP15)", "TEQUEX_CEQUIP15")
            .withColumnRenamed("sum(TEQUEX_CEQUIP16)", "TEQUEX_CEQUIP16")
            .withColumnRenamed("sum(TEQUEX_CEQUIP17)", "TEQUEX_CEQUIP17");
```

- Get only with the principal member of the family

```
tmiembDataset = tmiembDataset.filter(tmiembDataset.col("TMIEMB_CINDFA").equalTo("1"));
tusuinDataset = tusuinDataset.filter(tusuinDataset.col("TUSUIN_CINDFA").equalTo("1"));
```

- Join the values of different datasets by expedient code and intervention number: For each kind of problems defined in section C, a new dataset joining the values of original datasets is created

```
Dataset<Row> joinedTvalinDataset = tinterDataset
                .join(tvalinDataset,
                        tinterDataset.col("TINTER_CEXCOM").equalTo(tvalinDataset.col("
                            TVALIN_CEXCOM"))
                                .and(tinterDataset.col("TINTER_NINTER").equalTo(
                                    tvalinDataset.col("TVALIN_NINTER"))),
                        "inner")
                .join(tusuinDataset,
                        tinterDataset.col("TINTER_CEXCOM").equalTo(tusuinDataset.col("
                            TUSUIN_CEXCOM"))
```

```
                                    .and(tinterDataset.col("TINTER_NINTER").equalTo(
                                        tusuinDataset.col("TUSUIN_NINTER"))),
                            "inner")
                    .join(tmiembDataset,
                            tusuinDataset.col("TUSUIN_CEXCOM").equalTo(tmiembDataset.col("
                                TMIEMB_CEXCOM"))
                                    .and(tusuinDataset.col("TUSUIN_CINDFA").equalTo(
                                        tmiembDataset.col("TMIEMB_CINDFA"))),
                            "inner")
                    .join(texpfaDataset, tinterDataset.col("TINTER_CEXCOM").equalTo(
                        texpfaDataset.col("TEXPFA_CEXCOM")),
                            "inner")
                    .join(tExpfaEquexBinary,
                            tinterDataset.col("TINTER_CEXCOM").equalTo(tExpfaEquexBinary.col("
                                TEQUEX_CEXCOM")), "cross");
```

- Remove non-visible columns and reduced dates by year

```
for (FieldDTO tField : fields) {
  try {
      if ("reduce_date_year".equals(tField.getFieldTranslate())) {
        dataset = dataset.withColumn(colName(tField), dataset.col(colName(tField)).substr
            (0, 4));
      }
      if (!tField.getVisible()) {
        dataset = dataset.drop(dataset.col(colName(tField)));
      }
    } catch (Exception e) {
  }
}
```

- Save generated datasets in CSV files.

```
dataset.write().option("header", true).mode(SaveMode.Overwrite).csv(csvSavedFile.concat(
    name));
```

After execute this ETL service, developed with Apache Spark SQL, A set of CSV files for each of problems to be resolved has been obtained.

F. Machine Learning system

To develop the different algorithms of machine learning, platform H_2O.ai has been chosen, using Python as programing language.

Now, the different steps developed for each defined problems are enumerated

1) Build datasets:

- Get the list of fields: Invoke service siuss/data/field described in section D

```
url = "http://localhost:8081/siuss/data/field/visible"

response = requests.get(url)

fields = []
if (response.status_code == 200):
  fields = response.json()
else:
  print("ERROR! An error occurred trying to load fields")
    exit
```

- Load CSV files generates in ETL phase

```
data = h2o.import_file(path)
```

- Remove all columns where more than 80% of its values are nulls

```
cols_na = data.filter_na_cols(0.8)
data = data[cols_na]
```

- For each loaded field, the next steps is made:
  - If type of columns is ”code”, convert it to a factor, and if only exist a unique value for the category, drop it. Else, add this column to x
  - If type of columns is ”numeric” or ”date”, convert it to a numeric, show the histogram and add this column to x
  - If type of columns is ”label”, convert it to a factor and set y with the name of the column

```
y = ''
x = []

columns = data.col_names

# Categorize columns by field type
for field in fields:
  try:
    col = field['name']
    if col in columns:
      if field['fieldType'] == 'code':
        # Categorize factor columns
        print("Column '" + col + "' is a category")
                data[col] = data[col].asfactor()
                print("Categories for column: " + col)
                print(data[col].categories())
        if len(data[col].categories()) == 1:
                  print("Drop column " + col + " because has a unique category")
                  data = data.drop(col)
        else:
          x.append(col)
      elif field['fieldType'] == 'numeric':
              data[col] = data[col].asnumeric()
              x.append(col)
              # Show histogram for numeric columns
              print("Column '" + col + "' is a numeric")
              data[col].hist(plot=True)
      elif field['fieldType'] == 'date':
              data[col] = data[col].asnumeric()
              x.append(col)
              # Show histogram for numeric columns
              print("Column '" + col + "' is a date")
              data[col].hist(plot=True)
      elif field['fieldType'] == 'label':
              y = col
              # Simplify and categorize CODCOM columns
              print("Column '" + col + "' is the label")
              data[col] = data[col].asfactor()
              print("Categories for column: " + col)
              print(data[col].categories())
    else:
      print("Column '" + col + "' is NOT in dataset")
  except Exception as e:
    print("An error ocurred for column '" + col + "' -> " + str(e.__class__) + "." + str(e.
        __cause__))
```

2) Declaration and execution of machine learning model: In H_2O.ai, exists different kinds of machine learning models, that make very easy build and run it.

- Declare the machine learning model: Declare the model to run and define parameters of this model

```
m = H2ORandomForestEstimator(
        model_id="tvalin_maltrato_rf",
        seed=1234,
        ignore_const_cols=True,
        ntrees=100,
        stopping_metric="logloss",
        stopping_rounds=3,
        stopping_tolerance=0.02,
        max_runtime_secs=60,
        nfolds=10
        )
```

- Split the dataset to obtain train and test datasets

```
train, test = dataset.split_frame([0.8], seed=1234)
```

- Run built model with train dataset

```
m.train(x, y, train)
```

3) Validation of the model: After run the machine learning model, the performance of it is evaluated with our test dataset

```
performance = m.model_performance(test)
```

4) Get the MOJO of the model: Once a machine learning model has been found that behaves in an appropriate way for each of the problems to be addressed, a MOJO of it is downloaded. A MOJO (Model Object, Optimized) allows you to convert models that you build to Java jar application, which can then be deployed for scoring in real time.

```
modelfile = m.download_mojo(path=path + "/siuss-models/" + m.model_id, get_genmodel_jar=True)
```

In appendix D, the H_2O.ai model output is showed

G. Web application

Once we get our machine learning models for each of problems addressed, and we are stored it in H_2O.ai MOJOs, an application is developed to make real-time predictions for a new intervention.

1) Predict service: For that, a Spring Boot application will be developed that import downloaded H_2O.ai MOJOs. Each service receive as input the information of an intervention and returns the predictions made using the machine learning model stored in the MOJO.

Two Java Bean has been developed, one for the input and the other one for the result of the prediction:

- PredictInDTO
  - modelPath: Path where the downloaded MOJO has been stored
  - rows: List of input values. Each of value has a rowKey, that represent the name of the column in the dataset and a rowValue, that represent the assigned value of it
- PredictOutDTO
  - cod: Code of the label
  - description: Description of the label
  - prob: Probability that this label applies in our intervention
  - probPretty: Probability in #.## format

For each of problems addressed, a REST service has been developed to make the predictions for a new intervention

TABLE VIII
SIUSS EVALUATION PREDICT

Endpoint	**siuss/predict/tvalin**
Description	Makes predictions for evaluations
Type	POST
Input	oscuroweb.bigdata.dto.PredictInDTO
Output	oscuroweb.bigdata.dto.PredictOutDTO.

TABLE IX
SIUSS RECOURSE TO APPLY PREDICT

Endpoint	**siuss/predict/trecap**
Description	Makes predictions for resources to apply
Type	POST
Input	oscuroweb.bigdata.dto.PredictInDTO
Output	oscuroweb.bigdata.dto.PredictOutDTO.

TABLE X
SIUSS IDEAL RESOURCE PREDICT

Endpoint	**siuss/predict/trecid**
Description	Makes predictions for ideal resources
Type	POST
Input	oscuroweb.bigdata.dto.PredictInDTO
Output	oscuroweb.bigdata.dto.PredictOutDTO.

TABLE XI
SIUSS ABUSE PREDICT

Endpoint	**siuss/predict/maltrato**
Description	Makes prediction for if exist or not abuse in family unit
Type	POST
Input	oscuroweb.bigdata.dto.PredictInDTO
Output	oscuroweb.bigdata.dto.PredictOutDTO.

TABLE XII
SIUSS MATCHES PREDICT

Endpoint	**siuss/predict/coin**
Description	Makes predictions for if resources to apply matches ideal resources
Type	POST
Input	oscuroweb.bigdata.dto.PredictInDTO
Output	oscuroweb.bigdata.dto.PredictOutDTO.

Each REST service uses downloaded MOJOS to return prediction made for an input. The output is an ordered list of labels with it probability to apply in the intervention

2) Web application: To use all this system described in previous sections for social workers, a web application has been developed, implemented in Vaadin framework and Spring Boot.

This application consists in a input form, where the social worker will populate fields with the values of a new intervention and a results screen where all predictions made for each problems are showed.

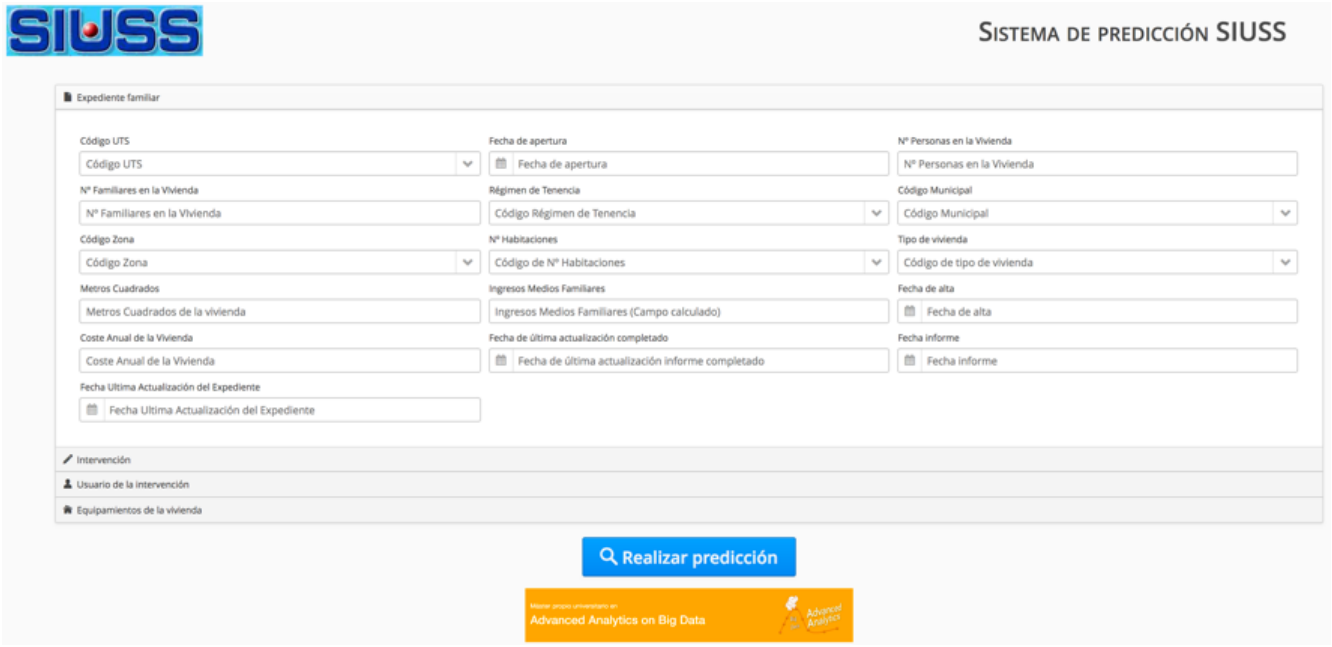


Fig. 2. SIUSS Web Application Form

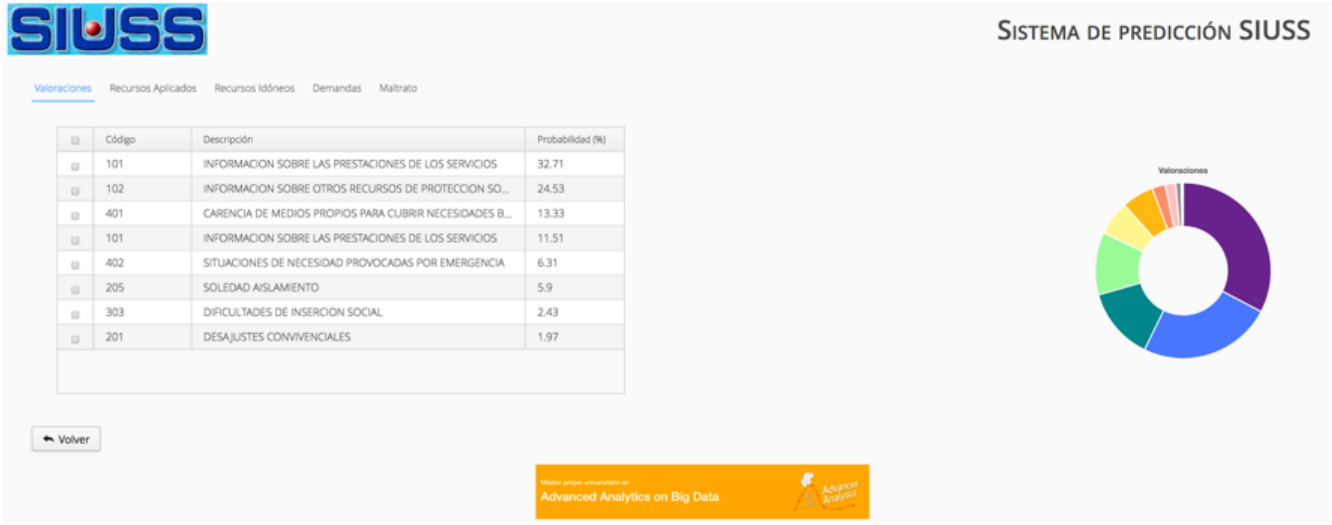


Fig. 3. SIUSS Web Application Results

In appendix E, a schema of the system is showed

IV. Experiments

A. Evaluations problems

Experiments made for predictions about evaluations.

There are 447 different evaluations. This may lead to difficulty to make reliable predictions. Fortunately, these evaluations have a hierarchy structure. For that, top levels of this hierarchy have been used, reducing the number of possible evaluations to 17.

The difficulty in addressing this problems is that a set of features will have different correct labels. Because of that, good performance is not expected.

1) Dataset:

- Size: 162527 x 61
- Type: Multinomial Classification
- Label categories: ['101000', '102000', '103000', '201000', '202000', '203000', '204000', '205000', '301000', '302000', '303000', '401000', '402000', '1010000']
- Label categories size: 17

2) Random Forest Model: A cartesian grid of parameters has been used to evaluate this problems with Random Forest models.

Parameters used:

- nfold: 10
- ignore_const_cols: True
- stopping_metric: "misclassification"
- stopping_rounds: 3
- stopping_tolerance: 0.02
- ntrees: [50, 75, 100] - grid parameter
- max_depth: [20, 40] - grid parameter
- min_rows: [1, 3, 5] - grid parameter

In appendix F there is a list of parameter used and its meaning.

These grid configuration generates 18 different models, one for each parameters combination.

3) Generalized Linear Models: A flexible generalization of ordinary linear regression.

Parameters used:

- nfold: 10
- ignore_const_cols: True
- family: ”multinomial”

In appendix F there is a list of parameter used and its meaning.

4) Deep Learning: Neural Nets.

Parameters used:

- ignore_const_cols: True
- epochs: 1000
- train_samples_per_iteration: 0
- stopping_metric: ”misclassification”
- stopping_rounds: 3
- stopping_tolerance: 0.02
- nfolds: 10

In appendix F there is a list of parameter used and its meaning.

B. Applied resources problems

Experiments made for predictions about applied resources.

There are 720 different resources. In the same way as prediction about evaluations, this may lead to difficulty to make reliable predictions. And as in evaluations, these resources have a hierarchy structure. For that, top levels of this hierarchy have been used, reducing the number of possible resources to 34.

The difficulty in addressing this problems is that a set of features will have different correct labels. For that, good performance is not expected.

1) Dataset:

- Size: 147534 x 62
- Type: Multinomial Classification
- Label categories: [’101000’, ’102000’, ’103000’, ’104000’, ’105000’, ’106000’, ’107000’, ’201000’, ’202000’, ’203000’, ’204000’, ’205000’, ’301000’, ’302000’, ’303000’, ’304000’, ’305000’, ’306000’, ’401000’, ’402000’, ’403000’, ’404000’, ’405000’, ’501000’, ’502000’, ’503000’, ’504000’, ’505000’, ’1010000’]
- Label categories size: 34

2) Random Forest Model: A cartesian grid of parameters has been used to evaluate this problems with Random Forest models.

Parameters used:

- nfold: 10
- ignore_const_cols: True
- stopping_metric: ”misclassification”
- stopping_rounds: 3
- stopping_tolerance: 0.02
- ntrees: [50, 75, 100] - grid parameter
- max_depth: [20, 40] - grid parameter
- min_rows: [1, 3, 5] - grid parameter

In appendix F there is a list of parameter used and its meaning.

These grid configuration generates 18 different models, one for each parameters combination.

3) Generalized Linear Models: A flexible generalization of ordinary linear regression.

Parameters used:

- nfold: 10
- ignore_const_cols: True
- family: ”multinomial”

In appendix F there is a list of parameter used and its meaning.

4) Deep Learning: Neural Nets.

Parameters used:

- ignore_const_cols: True
- epochs: 1000
- train_samples_per_iteration: 0
- stopping_metric: ”misclassification”
- stopping_rounds: 3
- stopping_tolerance: 0.02
- nfolds: 10

In appendix F there is a list of parameter used and its meaning.

C. Ideal resources problems

Experiments made for predictions about ideal resources.

There are 720 different resources. In the same way as prediction about evaluations, this may lead to difficulty to make reliable predictions. And as in evaluations, these resources have a hierarchy structure. For that, top levels of this hierarchy have been used, reducing the number of possible resources to 34.

The difficulty in addressing this problems is that a set of features will have different correct labels. For that, good performance is not expected.

1) Dataset:

- Size: 146721 x 61
- Type: Multinomial Classification
- Label categories: [’101000’, ’102000’, ’103000’, ’104000’, ’105000’, ’106000’, ’107000’, ’201000’, ’202000’, ’203000’, ’204000’, ’205000’, ’301000’, ’302000’, ’303000’, ’304000’, ’305000’, ’306000’, ’401000’, ’402000’, ’403000’, ’404000’, ’405000’, ’501000’, ’502000’, ’503000’, ’504000’, ’505000’, ’1010000’]
- Label categories size: 34

2) Random Forest Model: A cartesian grid of parameters has been used to evaluate this problems with Random Forest models.

Parameters used:

- nfold: 10
- ignore_const_cols: True
- stopping_metric: ”misclassification”
- stopping_rounds: 3
- stopping_tolerance: 0.02
- ntrees: [50, 75, 100] - grid parameter
- max_depth: [20, 40] - grid parameter
- min_rows: [1, 3, 5] - grid parameter

In appendix F there is a list of parameter used and its meaning.

These grid configuration generates 18 different models, one for each parameters combination.

3) Generalized Linear Models: A flexible generalization of ordinary linear regression.

Parameters used:

- nfold: 10
- ignore_const_cols: True
- family: ”multinomial”

In appendix F there is a list of parameter used and its meaning.

4) Deep Learning: Neural Nets.

Parameters used:

- ignore_const_cols: True
- epochs: 1000
- train_samples_per_iteration: 0
- stopping_metric: ”misclassification”
- stopping_rounds: 3
- stopping_tolerance: 0.02
- nfolds: 10

In appendix F there is a list of parameter used and its meaning.

D. Abuse problems

Experiments made for predictions about if exist abuse or not.

In contrast to before problems, only two categories are possible, it is a binomial problems, and a sets of features, only have one correct label. For that, this problem is simpler.

1) Dataset:

- Size: 105872 x 60
- Type: Binomial Classification
- Label categories: [’0’, ’1’]
- Label categories size: 2
- Label mean: 0.04

2) Random Forest Model: A cartesian grid of parameters has been used to evaluate this problems with Random Forest models.

Parameters used:

- nfold: 10
- ignore_const_cols: True
- stopping_metric: ”logloss”
- stopping_rounds: 3
- stopping_tolerance: 0.02
- fold_assignment: ”Stratified”
- balance_classes: True
- ntrees: [50, 75, 100] - grid parameter
- max_depth: [20, 40] - grid parameter
- min_rows: [1, 3, 5] - grid parameter

In appendix F there is a list of parameter used and its meaning.

These grid configuration generates 18 different models, one for each parameters combination.

E. Resources matching problems

Experiments made for predictions about if resource to apply matches ideal resource

This is, also, a binomial problems.

1) Dataset:

- Size: 105872 x 60
- Type: Binomial Classification
- Label categories: ['0', '1']
- Label categories size: 2
- Label mean: 0.9201

2) Random Forest Model: A cartesian grid of parameters has been used to evaluate this problems with Random Forest models.

Parameters used:

- nfold: 10
- ignore_const_cols: True
- stopping_metric: "logloss"
- stopping_rounds: 3
- stopping_tolerance: 0.02
- fold_assignment: "Stratified"
- balance_classes: True
- ntrees: [50, 75, 100] - grid parameter
- max_depth: [20, 40] - grid parameter
- min_rows: [1, 3, 5] - grid parameter

In appendix F there is a list of parameter used and its meaning.

These grid configuration generates 18 different models, one for each parameters combination.

V. Results

A. Evaluations Problems

1) Random Forest: In grid execution on Random Forest, this was the best combination of parameter:

max_depth	min_rows	ntrees	model_ids	logloss
20	5.0	8	tvalin_3_grid_model_4	1.8325590302063421
20	3.0	8	tvalin_3_grid_model_2	1.9249857416732215
20	1.0	8	tvalin_3_grid_model_0	2.0838875232769567
20	1.0	8	tvalin_3_grid_model_6	2.120909706349064
40	5.0	7	tvalin_3_grid_model_5	2.3109662393470547
40	3.0	8	tvalin_3_grid_model_3	2.788288848377348
40	1.0	7	tvalin_3_grid_model_1	4.019581914110473

Fig. 4. Evaluations Problems - Grid Results

For "tvalin_3_grid_model_4 models", this is the summary:

▾ OUTPUT - CROSS-VALIDATION METRICS SUMMARY

	mean	sd	cv_1_valid	cv_2_valid	cv_3_valid	cv_4_valid	cv_5_valid	cv_6_valid	cv_7_valid	cv_8_valid	cv_9_valid	cv_10_valid
accuracy	0.41673747	0.002258522	0.4190684	0.41635516	0.41231287	0.41927448	0.41073325	0.41950428	0.4209315	0.41405	0.4180791	0.41706565
err	0.5832625	0.002258522	0.5809316	0.5836448	0.58768713	0.5807255	0.5892667	0.5804957	0.57906854	0.58594996	0.5819209	0.5829344
err_count	7586.6	72.84174	7458.0	7651.0	7694.0	7460.0	7763.0	7518.0	7547.0	7707.0	7519.0	7549.0
logloss	1.8324251	0.027956048	1.7784116	1.8449523	1.8310007	1.8306733	1.864941	1.8152242	1.7734497	1.8638861	1.8102908	1.9114211
max_per_class_error	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
mean_per_class_accuracy	0.22265938	0.0032640512	0.2262816	0.21724884	0.21362613	0.22829929	0.21794227	0.22571543	0.22655568	0.22193907	0.22577044	0.22321507
mean_per_class_error	0.7773406	0.0032640512	0.7737184	0.78275114	0.78637385	0.7717007	0.7820577	0.7742846	0.77344435	0.7780609	0.7742295	0.7767849
mse	0.52278167	0.0015134631	0.52273226	0.5212154	0.5260199	0.52234143	0.52545124	0.5220657	0.5185811	0.5246728	0.5210504	0.52368635
r2	0.9752093	1.20956334E-4	0.9751108	0.97534937	0.97494555	0.97526556	0.9751541	0.975223	0.97547835	0.9749231	0.9753842	0.9752586
rmse	0.7230349	0.0010468608	0.72300225	0.72195244	0.72527224	0.7227319	0.72488016	0.72254115	0.7201258	0.724343	0.72183824	0.7236618

Fig. 5. Evaluations Problems - RF summary

▾ CROSS VALIDATION METRICS - CONFUSION MATRIX ROW LABELS: ACTUAL CLASS; COLUMN LABELS: PREDICTED CLASS

	101000	102000	103000	201000	202000	203000	204000	205000	301000	302000	303000	401000	402000	1010000	Error	Rate
101000	17575	623	1	349	1	24	1588	1716	1271	59	486	10314	73	76	**0.4854**	**16,581 / 34,156**
102000	2847	2198	0	260	0	16	182	197	607	22	669	3401	32	145	**0.7922**	**8,378 / 10,576**
103000	28	4	0	2	0	0	4	3	8	0	1	33	0	1	**1.0**	**84 / 84**
201000	1487	178	0	1366	2	56	610	254	714	42	187	2425	17	8	**0.8140**	**5,980 / 7,346**
202000	28	4	0	23	2	2	2	2	17	1	5	72	0	0	**0.9873**	**156 / 158**
203000	306	45	0	218	0	112	38	81	88	50	27	440	6	1	**0.9207**	**1,300 / 1,412**
204000	2534	50	0	187	0	8	4314	1206	36	0	64	480	7	54	**0.5174**	**4,626 / 8,940**
205000	4225	54	0	142	0	22	1802	1928	167	1	69	765	6	20	**0.7905**	**7,273 / 9,201**
301000	1866	296	2	477	2	12	173	200	2873	22	768	5732	37	1	**0.7694**	**9,588 / 12,461**
302000	340	32	0	116	0	22	0	2	64	188	27	296	1	1	**0.8274**	**901 / 1,089**
303000	1015	657	0	232	1	16	246	133	765	5	1532	1672	16	14	**0.7570**	**4,772 / 6,304**
401000	7332	481	0	690	3	24	601	376	2326	23	534	21940	95	12	**0.3629**	**12,497 / 34,437**
402000	813	49	0	38	0	2	75	42	143	3	32	1693	61	0	**0.9793**	**2,890 / 2,951**
1010000	472	126	0	5	0	0	115	69	2	0	21	30	0	112	**0.8824**	**840 / 952**
Total	**40868**	**4797**	**3**	**4105**	**11**	**316**	**9750**	**6209**	**9081**	**416**	**4422**	**49293**	**351**	**445**	**0.5833**	**75,866 / 130,067**

Fig. 6. Evaluations Problems - RF confusion matrix

2) Generalized Linear Models: This is the summary for GLM execution

▾ OUTPUT - CROSS-VALIDATION METRICS SUMMARY

	mean	sd	cv_1_valid	cv_2_valid	cv_3_valid	cv_4_valid	cv_5_valid	cv_6_valid	cv_7_valid	cv_8_valid	cv_9_valid	cv_10_valid
accuracy	0.38135883	0.0026562917	0.3805471	0.3827491	0.3833989	0.3772349	0.38293484	0.37833863	0.38907504	0.37519297	0.3839577	0.38015896
err	0.6186412	0.0026562917	0.6194529	0.61725086	0.6166011	0.6227651	0.61706513	0.62166137	0.61092496	0.62480706	0.61604226	0.61984104
err_count	8037.0	63.021423	8152.0	8065.0	7993.0	8081.0	8078.0	8030.0	7840.0	8095.0	7926.0	8110.0
logloss	1.6899157	0.004877187	1.6935765	1.6995595	1.6851045	1.6868176	1.6913829	1.6911244	1.6754376	1.6949786	1.6975454	1.6836296
max_per_class_error	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
mean_per_class_accuracy	0.15854877	0.0033417845	0.16151735	0.15902017	0.1669826	0.15924522	0.15621485	0.15065631	0.16308203	0.1519997	0.15599015	0.16077939
mean_per_class_error	0.8414512	0.0033417845	0.8384826	0.8409798	0.8330174	0.8407548	0.84378517	0.8493437	0.836918	0.8480003	0.8440098	0.8392206
mse	0.5767012	0.002054125	0.57655793	0.57688415	0.58223164	0.57650214	0.5743811	0.5752309	0.5727915	0.5764891	0.57425654	0.581687
null_deviance	52981.605	331.27942	53692.37	53520.617	52787.117	52965.74	53478.21	52391.152	52235.555	52901.223	52609.21	53234.87
r2	0.97264975	1.4368647E-4	0.97273874	0.972524	0.9724419	0.9725188	0.9726423	0.97289765	0.9729262	0.97281927	0.9727327	0.97225565
residual_deviance	43908.496	304.13095	44574.934	44412.887	43688.02	43776.293	44283.785	43688.508	43001.78	43920.285	43681.24	44057.223
rmse	0.75940603	0.0013510523	0.7593141	0.7595289	0.763041	0.7592774	0.7578794	0.7584398	0.7568299	0.75926876	0.7577972	0.76268405

Fig. 7. Evaluations Problems - GLM summary

▾ CROSS VALIDATION METRICS - CONFUSION MATRIX ROW LABELS: ACTUAL CLASS; COLUMN LABELS: PREDICTED CLASS

	101000	102000	103000	201000	202000	203000	204000	205000	301000	302000	303000	401000	402000	1010000	Error	Rate
101000	17203	1	0	365	0	0	1785	1390	865	34	162	12265	0	0	**0.4951**	**16,867 / 34,070**
102000	3367	33	0	310	0	0	343	187	547	9	587	5259	1	0	**0.9969**	**10,610 / 10,643**
103000	40	0	0	6	0	0	1	1	0	0	0	39	0	0	**1.0**	**87 / 87**
201000	1787	0	0	730	0	0	591	190	676	14	32	3363	0	0	**0.9011**	**6,653 / 7,383**
202000	31	0	0	24	0	0	2	0	6	0	1	100	0	0	**1.0**	**164 / 164**
203000	424	0	0	169	0	0	45	40	160	15	2	543	0	0	**1.0**	**1,398 / 1,398**
204000	3623	0	0	27	0	0	4201	755	24	1	3	337	0	0	**0.5317**	**4,770 / 8,971**
205000	5068	0	0	41	0	0	1782	1406	90	0	9	826	0	0	**0.8475**	**7,816 / 9,222**
301000	1894	0	0	467	0	0	175	114	1739	5	327	7737	0	0	**0.8604**	**10,719 / 12,458**
302000	439	0	0	157	0	0	0	0	49	37	5	394	0	0	**0.9658**	**1,044 / 1,081**
303000	1238	0	0	144	0	0	234	145	529	2	842	3030	0	0	**0.8634**	**5,322 / 6,164**
401000	7848	0	0	501	0	0	750	321	1291	9	167	23351	1	0	**0.3180**	**10,888 / 34,239**
402000	816	0	0	47	0	0	90	30	57	2	4	2016	0	0	**1.0**	**3,062 / 3,062**
1010000	514	0	0	5	0	0	275	147	4	0	1	24	0	0	**1.0**	**970 / 970**
Total	**44292**	**34**	**0**	**2993**	**0**	**0**	**10274**	**4726**	**6037**	**128**	**2142**	**59284**	**2**	**0**	**0.6186**	**80,370 / 129,912**

Fig. 8. Evaluations Problems - GLM confusion matrix

3) Deep Learning - Neural Nets: This is the summary for Deep Learning execution

▾ OUTPUT - CROSS-VALIDATION METRICS SUMMARY

	mean	sd	cv_1_valid	cv_2_valid	cv_3_valid	cv_4_valid	cv_5_valid	cv_6_valid	cv_7_valid	cv_8_valid	cv_9_valid	cv_10_valid
accuracy	0.42821795	0.0034855856	0.4226583	0.42783192	0.43736145	0.43302205	0.42512375	0.43164352	0.4216533	0.42598608	0.43267754	0.42422146
err	0.57178205	0.0034855856	0.57734174	0.5721681	0.5626385	0.566978	0.57487625	0.5683565	0.5783467	0.57401395	0.5673225	0.57577854
err_count	7436.1	47.557808	7532.0	7516.0	7361.0	7407.0	7432.0	7321.0	7500.0	7422.0	7382.0	7488.0
logloss	1.5616826	0.007904761	1.5663593	1.5596542	1.5532839	1.5489092	1.5708064	1.5615194	1.5654476	1.58152	1.5688138	1.5405118
max_per_class_error	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
mean_per_class_accuracy	0.22452697	0.0068717655	0.22667637	0.22701569	0.23568909	0.23369654	0.20593871	0.23250522	0.20773314	0.22973943	0.22429925	0.22197634
mean_per_class_error	0.775473	0.0068717655	0.77332366	0.7729843	0.7643109	0.7663035	0.7940613	0.7674948	0.79226685	0.7702606	0.77570075	0.77802366
mse	0.5147312	0.0025651178	0.5154545	0.51976556	0.5066817	0.51202756	0.51754254	0.5167392	0.5180926	0.51158863	0.5142033	0.51521695
r2	0.9755871	1.7113429E-4	0.9755452	0.9750274	0.97599703	0.97578883	0.97560316	0.9756657	0.9753943	0.97562945	0.9756982	0.9755219
rmse	0.7174433	0.0017901879	0.7179516	0.7209477	0.7118158	0.715561	0.7194043	0.7188457	0.71978647	0.71525425	0.7170797	0.71778613

Fig. 9. Evaluations Problems - Deep Learning summary

▾ CROSS VALIDATION METRICS - CONFUSION MATRIX ROW LABELS: ACTUAL CLASS; COLUMN LABELS: PREDICTED CLASS

	101000	102000	103000	201000	202000	203000	204000	205000	301000	302000	303000	401000	402000	1010000	Error	Rate
101000	17553	407	0	305	0	22	1869	1212	1093	110	508	10867	57	89	**0.4851**	**16,539 / 34,092**
102000	2667	2034	0	213	0	25	238	127	617	35	521	4020	16	101	**0.8084**	**8,580 / 10,614**
103000	37	4	0	1	0	0	2	4	5	0	3	34	0	0	**1.0**	**90 / 90**
201000	1473	135	0	1057	0	49	652	148	760	54	131	2886	10	21	**0.8567**	**6,319 / 7,376**
202000	25	4	0	19	0	0	2	0	20	3	2	80	0	0	**1.0**	**155 / 155**
203000	341	33	0	187	0	85	48	62	147	46	20	454	0	1	**0.9403**	**1,339 / 1,424**
204000	2600	27	0	84	0	7	4933	648	46	0	53	520	5	66	**0.4512**	**4,056 / 8,989**
205000	4219	40	0	78	0	29	2121	1600	156	3	73	836	1	33	**0.8259**	**7,589 / 9,189**
301000	1702	228	0	422	0	13	234	130	2717	29	660	6216	36	0	**0.7807**	**9,670 / 12,387**
302000	327	29	0	88	0	18	1	1	86	212	14	341	1	0	**0.8104**	**906 / 1,118**
303000	899	562	0	162	0	10	301	89	618	10	1709	1915	4	15	**0.7285**	**4,585 / 6,294**
401000	6493	375	0	552	0	24	755	262	1714	34	436	23618	73	11	**0.3124**	**10,729 / 34,347**
402000	707	32	0	44	0	5	90	36	96	1	33	1925	55	3	**0.9818**	**2,972 / 3,027**
1010000	542	46	0	5	0	0	138	43	0	1	16	41	0	119	**0.8749**	**832 / 951**
Total	**39585**	**3956**	**0**	**3217**	**0**	**287**	**11384**	**4362**	**8075**	**538**	**4179**	**53753**	**258**	**459**	**0.5718**	**74,361 / 130,053**

Fig. 10. Evaluations Problems - Deep Learning confusion matrix

B. Applied resources problems

1) Random Forest: In grid execution on Random Forest, this was the best combination of parameter:

max_depth	min_rows	ntrees	model_ids	logloss
20	5.0	9	trecap_3_grid_model_4	2.1692408313049008
20	3.0	10	trecap_3_grid_model_2	2.2417336945294264
20	1.0	9	trecap_3_grid_model_0	2.3699331389916236
40	5.0	9	trecap_3_grid_model_5	2.637829817362878
40	3.0	9	trecap_3_grid_model_3	3.1255258246450817
40	1.0	9	trecap_3_grid_model_1	4.1869644446672893

Fig. 11. Applied Resources Problems - Grid Results

For ”trecap_3_grid_model_4 models”, this is the summary:

▾ OUTPUT - CROSS-VALIDATION METRICS SUMMARY

	mean	sd	cv_1_valid	cv_2_valid	cv_3_valid	cv_4_valid	cv_5_valid	cv_6_valid	cv_7_valid	cv_8_valid	cv_9_valid	cv_10_valid
accuracy	0.47515577	0.0031680395	0.47657046	0.4732538	0.4762749	0.47730362	0.46494836	0.47968856	0.47762987	0.47892785	0.4778701	0.46909028
err	0.5248442	0.0031680395	0.5234295	0.5267462	0.5237251	0.5226964	0.53505164	0.5203114	0.5223701	0.52107215	0.5221299	0.5309097
err_count	6196.6	56.86317	6166.0	6312.0	6203.0	6149.0	6373.0	6148.0	6083.0	6182.0	6158.0	6192.0
logloss	2.1863892	0.06950953	2.187862	2.2855864	2.0377157	2.1645858	2.3235204	2.154151	2.1441977	2.0304754	2.2179546	2.317843
max_per_class_error	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
mean_per_class_accuracy	0.16742915	0.007467365	0.1644429	0.16424984	0.16064617	0.17318068	0.15907176	0.1958714	0.1568677	0.16680717	0.17010719	0.16304667
mean_per_class_error	0.83257085	0.007467365	0.8355571	0.83575016	0.83935386	0.8268193	0.84092826	0.8041286	0.8431323	0.8331928	0.8298928	0.83695334
mse	0.52360624	0.002263974	0.52382797	0.52175266	0.5239709	0.518842	0.5308593	0.52137077	0.52097464	0.52487123	0.522895	0.5266981
r2	0.9945899	4.9511465E-5	0.9945697	0.9945315	0.9946317	0.9946854	0.994464	0.99457806	0.9946914	0.994574	0.9945245	0.9946492
rmse	0.72360307	0.0015624358	0.7237596	0.7223245	0.72385836	0.7203069	0.7286009	0.7220601	0.7217857	0.72448	0.7231148	0.7257397

Fig. 12. Applied Resources Problems - RF summary

▾ CROSS VALIDATION METRICS - CONFUSION MATRIX ROW LABELS: ACTUAL CLASS; COLUMN LABELS: PREDICTED CLASS

	101000	102000	103000	104000	105000	106000	107000	201000	202000	203000	204000	205000	301000	302000	303000	304000	305000	306000	401000	402000	403000	404000	405000	501000	502000	503000	504000	505000	1010000	Error	Rate
101000	28568	31	31	1685	504	82	1	2329	92	204	51	0	35	0	0	0	0	0	59	4	17	291	0	14	102	3549	17	10	114	0.3896	9,222 / 29,790
102000	942	107	1	88	10	6	0	76	2	12	1	0	1	0	0	0	0	0	0	0	0	19	0	0	5	476	1	1	2	0.9389	1,643 / 1,750
103000	536	4	47	596	39	37	0	141	3	17	8	0	15	0	0	0	0	0	19	2	1	84	0	1	19	428	1	6	38	0.9770	1,995 / 2,042
104000	3351	25	64	9466	561	315	2	535	11	77	89	0	60	0	1	1	0	0	24	3	24	657	1	5	48	1202	14	10	157	0.4333	7,237 / 16,703
105000	1057	4	6	1080	2226	46	1	1871	32	16	3	1	49	0	0	0	0	0	6	2	6	127	0	6	11	192	2	2	22	0.6711	4,542 / 6,768
106000	526	5	11	1515	59	423	0	51	0	17	0	0	5	0	0	0	1	0	5	3	0	181	0	1	24	230	16	1	0	0.8624	2,651 / 3,074
107000	48	0	0	71	6	5	1	9	0	1	0	0	1	0	0	0	0	0	0	0	0	13	0	0	0	20	0	0	0	0.9943	174 / 175
201000	1869	0	2	165	969	2	0	10101	203	7	1	0	32	0	0	0	0	0	4	0	0	7	0	8	1	409	0	3	28	0.2686	3,710 / 13,811
202000	441	0	0	58	265	0	0	1870	132	2	0	0	14	0	0	0	0	0	0	0	2	3	0	1	2	87	0	0	17	0.9544	2,762 / 2,894
203000	1615	7	5	227	16	15	0	95	0	639	9	0	1	0	0	0	0	0	22	2	8	140	0	0	47	579	0	3	1	0.8138	2,792 / 3,431
204000	251	0	3	302	70	4	0	103	3	12	70	0	4	0	0	0	0	0	4	1	0	39	0	1	15	194	1	1	0	0.9351	1,008 / 1,078
205000	24	0	0	10	3	0	0	29	1	0	1	0	1	0	0	0	0	0	1	0	0	1	0	0	1	24	0	0	2	1.0	98 / 98
301000	188	0	2	113	183	2	0	604	7	1	2	0	190	0	0	0	0	0	7	0	0	12	0	7	4	59	1	1	8	0.8634	1,201 / 1,391
302000	4	0	1	18	3	2	0	6	0	0	0	0	0	2	0	0	0	0	0	0	0	5	0	0	0	7	0	0	0	0.9583	46 / 48
303000	8	0	0	10	7	0	0	4	0	0	0	0	5	1	0	0	0	0	1	0	0	6	0	0	1	9	0	0	1	1.0	53 / 53
304000	14	0	0	13	3	0	0	3	0	0	0	0	2	0	0	0	0	0	0	0	0	1	0	1	1	6	0	0	0	1.0	44 / 44
305000	5	0	0	7	2	2	0	5	0	0	0	0	3	0	0	0	0	0	0	0	0	1	0	0	0	19	0	0	0	1.0	44 / 44
306000	9	0	0	9	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	4	10	0	0	0	1.0	38 / 38
401000	826	2	13	387	7	21	0	28	0	51	6	0	8	0	0	0	0	0	65	2	0	56	0	0	39	309	0	0	0	0.9643	1,755 / 1,820
402000	34	0	3	119	3	6	0	6	0	4	0	0	0	0	0	0	0	0	2	6	1	42	0	0	4	65	0	0	0	0.9797	289 / 295
403000	146	1	0	78	6	2	0	33	5	11	1	1	0	0	0	0	0	0	1	0	104	42	0	0	3	82	1	0	0	0.7988	413 / 517
404000	1098	7	12	1816	49	119	0	45	0	94	22	0	1	0	0	0	0	0	19	6	13	1452	0	1	67	799	15	4	0	0.7425	4,187 / 5,639
405000	10	0	0	1	4	0	0	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1.0	57 / 57
501000	88	0	0	47	43	3	0	44	1	2	0	0	2	0	0	0	0	0	1	0	0	6	0	146	3	64	0	0	3	0.6777	307 / 453
502000	1259	9	7	330	42	34	0	33	0	105	12	0	7	0	0	0	0	0	30	0	1	167	0	0	199	1030	2	9	0	0.9393	3,077 / 3,276
503000	7977	43	19	562	43	17	1	722	14	113	18	0	14	1	0	0	0	1	13	1	4	159	0	15	112	9804	24	21	19	0.5028	9,913 / 19,717
504000	367	4	2	240	11	38	0	28	2	6	0	0	0	0	0	0	0	0	0	0	0	51	0	0	8	361	23	6	1	0.9800	1,125 / 1,148
505000	264	5	1	78	1	7	0	32	2	10	0	0	1	0	0	0	0	0	0	1	1	19	0	1	19	356	3	28	2	0.9663	803 / 831
1010000	395	1	7	140	20	0	0	174	9	0	0	0	5	0	0	0	0	0	0	0	0	1	0	1	0	26	0	1	299	0.7229	780 / 1,079
Total	43920	255	237	19231	5155	1188	6	19021	519	1401	294	2	456	4	1	1	1	1	283	33	182	3584	1	209	739	20397	121	107	715	0.5249	61,966 / 118,064

Fig. 13. Applied Resources Problems - RF confusion matrix

2) Generalized Linear Models: This is the summary for GLM execution

▾ OUTPUT - CROSS-VALIDATION METRICS SUMMARY

	mean	sd	cv_1_valid	cv_2_valid	cv_3_valid	cv_4_valid	cv_5_valid	cv_6_valid	cv_7_valid	cv_8_valid	cv_9_valid	cv_10_valid
accuracy	0.44357237	0.002092765	0.43920946	0.4403999	0.44853875	0.44498548	0.44482145	0.44526404	0.44186643	0.43996626	0.44696006	0.44371197
err	0.5564276	0.002092765	0.56079054	0.5596001	0.5514612	0.55501455	0.5551786	0.55473596	0.55813354	0.56003374	0.55303997	0.556288
err_count	6570.9	48.2436	6725.0	6549.0	6510.0	6497.0	6545.0	6618.0	6519.0	6642.0	6522.0	6582.0
logloss	1.7702576	0.006147649	1.7794511	1.7672786	1.7542716	1.7760285	1.7760913	1.7609781	1.7686701	1.7783566	1.7803407	1.7611088
max_per_class_error	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
mean_per_class_accuracy	0.12153654	0.009332754	0.113555916	0.11420639	0.115097396	0.11930143	0.14397581	0.113118276	0.113134965	0.11497322	0.11698114	0.15102082
mean_per_class_error	0.87846345	0.009332754	0.8864441	0.8857936	0.8849026	0.88069856	0.8560242	0.8868817	0.886865	0.8850268	0.88301885	0.8489792
mse	0.56701803	0.0011163299	0.5682434	0.5680853	0.563334	0.5678286	0.5677418	0.56554717	0.5671082	0.5688672	0.5676486	0.5657764
null_deviance	56259.793	336.28027	57416.15	55741.26	56118.527	55911.277	55910.703	56508.656	55761.176	56505.27	56378.01	56346.91
r2	0.99412686	2.990317E-5	0.99414134	0.99409074	0.994206	0.99413866	0.99403954	0.994147	0.9941264	0.9941464	0.99414194	0.99409056
residual_deviance	41810.04	286.29742	42678.355	41364.92	41418.35	41580.38	41876.68	42016.938	41316.133	42182.617	41991.113	41674.88
rmse	0.75300527	7.418359E-4	0.7538192	0.7537143	0.75055575	0.75354403	0.75348645	0.7520287	0.7530659	0.7542328	0.7534246	0.7521811

Fig. 14. Applied Resources Problems - GLM summary

▾ CROSS VALIDATION METRICS - CONFUSION MATRIX ROW LABELS: ACTUAL CLASS; COLUMN LABELS: PREDICTED CLASS

	101000	102000	103000	104000	105000	106000	107000	201000	202000	203000	204000	205000	301000	302000	303000	304000	305000	306000	401000	402000	403000	404000	405000	501000	502000	503000	504000	505000	1010000	Error	Rate
101000	20967	0	0	1412	625	30	0	2667	20	108	2	0	12	0	0	0	0	0	0	0	0	157	0	10	4	3725	0	0	0	0.2950	8,772 / 29,739
102000	1005	1	0	57	7	0	0	74	0	6	0	0	2	0	0	0	0	0	0	0	0	11	0	0	2	587	0	0	0	0.9994	1,751 / 1,752
103000	576	0	0	580	105	17	0	150	3	7	2	0	13	0	0	0	0	0	0	0	1	78	0	0	1	526	0	0	2	1.0	2,069 / 2,069
104000	4324	0	1	8726	880	153	0	474	3	48	4	0	26	0	0	0	0	0	0	0	2	513	0	3	4	1581	0	0	4	0.4789	8,020 / 16,746
105000	1197	0	0	1090	2433	24	0	1657	3	7	0	0	15	0	0	0	0	0	0	0	0	159	0	3	1	239	0	0	2	0.6438	4,397 / 6,830
106000	450	0	0	1922	76	175	0	46	1	14	0	0	0	0	0	0	0	0	0	0	0	129	0	0	2	271	0	0	0	0.9433	2,911 / 3,086
107000	45	0	0	65	9	5	0	10	0	0	0	0	2	0	0	0	0	0	0	0	0	6	0	0	0	33	0	0	0	1.0	175 / 175
201000	2736	0	0	134	1653	1	0	8717	53	6	0	0	16	0	0	0	0	0	0	0	0	3	0	5	1	510	0	0	3	0.3701	5,121 / 13,838
202000	470	0	0	38	517	1	0	1748	46	1	0	0	5	0	0	0	0	0	0	0	0	11	0	0	0	84	0	0	2	0.9843	2,877 / 2,923
203000	2089	0	0	253	7	3	0	78	0	222	1	0	0	0	0	0	0	0	0	0	1	55	0	0	3	717	0	0	1	0.9353	3,208 / 3,430
204000	295	0	0	304	121	1	0	45	0	22	7	0	3	0	0	0	0	0	0	0	0	23	0	0	5	249	0	0	0	0.9935	1,068 / 1,075
205000	21	0	0	12	5	0	0	19	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	25	0	0	0	1.0	85 / 85
301000	213	0	0	70	431	0	0	491	5	0	1	0	93	0	0	0	0	0	0	0	0	9	0	5	1	70	0	0	3	0.9332	1,299 / 1,392
302000	2	0	0	30	3	0	0	3	0	0	0	0	1	0	0	0	0	0	0	0	0	3	0	1	0	4	0	0	0	1.0	47 / 47
303000	6	0	0	13	3	0	0	5	0	1	0	0	3	0	0	0	0	0	0	0	0	10	0	0	1	10	0	0	0	1.0	52 / 52
304000	7	0	0	11	5	0	0	4	0	1	0	0	1	0	0	0	0	0	0	0	0	3	0	0	0	12	0	0	0	1.0	44 / 44
305000	9	0	0	9	2	0	0	4	0	0	0	0	3	0	0	0	0	0	0	0	0	1	0	0	0	14	0	0	0	1.0	42 / 42
306000	7	0	0	18	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	11	0	0	0	1.0	40 / 40
401000	778	0	0	405	8	31	0	24	0	21	0	0	13	0	0	0	0	0	0	0	0	75	0	0	2	434	0	0	0	1.0	1,791 / 1,791
402000	27	0	0	133	6	2	0	10	0	1	0	0	0	0	0	0	0	0	0	0	0	40	0	0	0	71	0	0	0	1.0	290 / 290
403000	271	0	0	72	14	2	0	51	4	2	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	82	0	0	0	0.9825	506 / 515
404000	1449	0	0	2384	64	59	0	40	1	83	19	0	0	0	0	0	0	0	0	0	0	616	0	2	9	991	0	0	1	0.8923	5,102 / 5,718
405000	16	0	0	1	14	0	0	22	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1.0	56 / 56
501000	104	0	0	45	86	4	0	29	0	0	0	0	4	0	0	0	0	0	0	0	0	6	0	107	0	73	0	0	0	0.7664	351 / 458
502000	1263	1	0	449	42	5	0	36	1	40	2	0	5	0	0	0	0	0	0	0	0	135	0	0	43	1223	0	0	0	0.9867	3,202 / 3,245
503000	8035	0	0	362	32	2	0	788	6	68	1	0	4	0	0	0	0	0	0	0	1	60	0	5	11	10214	0	0	1	0.4786	9,376 / 19,590
504000	487	0	0	323	15	16	0	28	0	1	0	0	0	0	0	0	0	0	0	0	0	18	0	1	0	324	0	0	0	1.0	1,133 / 1,133
505000	274	0	0	92	2	0	0	27	0	0	0	0	2	0	0	0	0	0	0	0	0	10	0	0	0	437	0	0	1	1.0	845 / 845
1010000	605	0	0	4	9	0	0	427	0	0	0	0	2	0	0	0	0	0	0	0	0	3	0	0	0	29	0	0	5	0.9954	1,079 / 1,084
Total	47648	2	1	19022	7174	531	0	17676	146	660	39	0	226	0	0	0	0	0	0	0	14	2145	0	142	91	22548	0	0	25	0.5564	65,789 / 118,090

Fig. 15. Applied Resources Problems - GLM confusion matrix

3) Deep Learning - Neural Nets: This is the summary for Deep Learning execution

▾ OUTPUT - CROSS-VALIDATION METRICS SUMMARY

	mean	sd	cv_1_valid	cv_2_valid	cv_3_valid	cv_4_valid	cv_5_valid	cv_6_valid	cv_7_valid	cv_8_valid	cv_9_valid	cv_10_valid
accuracy	0.48105767	0.0030518142	0.48115298	0.4831739	0.48154762	0.47112462	0.48342636	0.48582962	0.48320457	0.47470915	0.48277327	0.4836345
err	0.51894236	0.0030518142	0.518847	0.5168261	0.5184524	0.52887535	0.51657367	0.5141704	0.51679546	0.52529085	0.5172267	0.51636547
err_count	6124.2	48.074734	6084.0	6051.0	6097.0	6264.0	6109.0	6114.0	6154.0	6231.0	6080.0	6058.0
logloss	1.6642805	0.013511031	1.6755161	1.6624606	1.6865376	1.6906614	1.6387663	1.6573404	1.677594	1.6721905	1.653013	1.6287255
max_per_class_error	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
mean_per_class_accuracy	0.17530194	0.0061414456	0.17401685	0.17022502	0.18912537	0.17257647	0.19045189	0.17847194	0.166539	0.16305368	0.16880529	0.17975387
mean_per_class_error	0.8246981	0.0061414456	0.82598317	0.829775	0.81087464	0.8274235	0.80954814	0.8215281	0.833461	0.8369463	0.8311947	0.8202461
mse	0.4959977	0.0034466027	0.5018284	0.49003008	0.49444634	0.5067757	0.49088052	0.49563658	0.49853083	0.49230343	0.49467355	0.49487174
r2	0.99487555	3.5951347E-5	0.9948057	0.99493897	0.994831	0.99477625	0.99491996	0.99488866	0.994889	0.99490345	0.9948827	0.99491966
rmse	0.7042626	0.0024415115	0.70839846	0.7000215	0.7031688	0.71188176	0.70062864	0.7040146	0.70606714	0.7016434	0.70333034	0.7034712

Fig. 16. Applied Resources Problems - Deep Learning summary

▾ CROSS VALIDATION METRICS - CONFUSION MATRIX ROW LABELS: ACTUAL CLASS; COLUMN LABELS: PREDICTED CLASS

	101000	102000	103000	104000	105000	106000	107000	201000	202000	203000	204000	205000	301000	302000	303000	304000	305000	306000	401000	402000	403000	404000	405000	501000	502000	503000	504000	505000	1010000	Error	Rate
101000	21065	77	29	1728	574	78	0	2069	130	382	65	0	73	0	0	0	0	0	37	0	24	292	0	34	87	2889	16	0	142	0.2929	8,726 / 29,791
102000	943	132	0	95	12	2	0	73	2	25	1	0	4	0	0	0	0	0	2	0	0	23	0	0	6	424	1	1	3	0.9245	1,617 / 1,749
103000	544	4	52	586	60	49	0	123	2	15	18	0	39	0	0	0	0	0	18	0	2	86	0	5	31	388	0	1	28	0.9746	1,999 / 2,051
104000	3263	25	34	9540	609	341	0	523	8	127	120	0	188	0	0	0	0	0	38	0	40	563	0	13	76	1189	10	3	99	0.4297	7,189 / 16,729
105000	990	4	5	1045	2531	64	0	1592	34	15	8	0	63	0	0	0	0	0	2	1	6	134	0	18	16	220	0	0	33	0.6268	4,250 / 6,781
106000	449	1	9	1654	51	411	0	34	0	24	3	0	21	0	0	0	0	0	10	0	4	174	0	0	18	238	2	1	0	0.8676	2,693 / 3,104
107000	45	0	0	70	5	5	0	16	0	3	4	0	4	0	0	0	0	0	0	0	0	8	0	0	0	18	0	0	0	1.0	178 / 178
201000	2206	4	1	187	1134	4	0	9282	260	18	1	0	49	0	0	0	0	0	2	0	1	18	0	13	2	426	1	0	79	0.3219	4,406 / 13,688
202000	413	0	0	51	332	0	0	1659	269	3	1	0	38	0	0	0	0	0	0	0	2	14	0	5	1	101	0	0	15	0.9074	2,635 / 2,904
203000	1570	6	3	233	10	21	0	86	1	704	15	0	1	0	0	0	0	0	12	0	13	145	0	1	29	576	2	0	4	0.7949	2,728 / 3,432
204000	196	2	1	243	77	7	0	101	0	18	162	0	4	0	0	0	0	0	3	0	0	50	0	1	13	191	0	0	1	0.8486	908 / 1,070
205000	21	1	0	19	2	0	0	22	0	3	0	0	2	0	0	0	0	0	0	0	1	2	0	0	2	17	0	0	2	1.0	94 / 94
301000	227	0	1	103	222	2	0	454	17	1	2	0	288	0	0	0	0	0	3	0	0	10	0	6	3	64	0	0	12	0.7965	1,127 / 1,415
302000	8	0	0	21	3	2	0	5	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	5	0	1	0	1.0	47 / 47
303000	7	0	1	7	7	1	0	4	0	0	0	0	6	0	0	0	0	0	0	0	0	9	0	0	0	6	0	0	0	1.0	48 / 48
304000	11	0	0	13	6	1	0	4	1	2	0	0	3	0	0	0	0	0	0	0	0	2	0	0	0	6	0	0	0	1.0	49 / 49
305000	6	0	0	9	3	0	0	3	0	0	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	15	0	0	0	1.0	39 / 39
306000	4	0	1	6	3	1	0	5	0	1	1	0	0	0	0	0	0	0	0	0	0	2	0	0	4	11	0	0	0	1.0	39 / 39
401000	708	2	13	355	9	35	0	30	1	83	11	0	18	0	0	0	0	0	71	1	1	73	0	2	31	357	0	1	0	0.9686	1,731 / 1,802
402000	25	0	4	117	4	13	0	2	0	1	0	0	3	0	0	0	0	0	3	1	1	50	0	0	2	68	0	0	0	0.9966	293 / 294
403000	125	1	0	83	10	4	0	26	11	11	4	0	1	0	0	0	0	0	1	0	132	30	0	1	4	70	0	0	1	0.7437	383 / 515
404000	975	8	13	2055	66	132	0	32	0	155	25	0	4	0	0	0	0	0	21	0	25	1296	0	3	77	787	9	1	1	0.7720	4,389 / 5,685
405000	14	0	0	0	16	0	0	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1.0	55 / 55
501000	78	0	2	56	43	2	0	26	4	1	0	0	11	0	0	0	0	0	0	0	0	8	0	172	3	52	0	0	2	0.6261	288 / 460
502000	1133	20	7	331	39	58	0	38	1	120	9	0	13	0	0	0	0	0	22	2	5	144	0	0	266	1055	1	2	0	0.9186	3,000 / 3,266
503000	7433	87	16	514	45	35	0	665	16	233	44	0	20	0	0	0	0	0	13	4	13	167	0	20	153	10102	28	5	34	0.4858	9,545 / 19,647
504000	383	7	0	290	12	31	0	35	1	5	1	0	1	0	0	0	0	0	2	0	1	41	0	1	4	341	7	0	1	0.9940	1,157 / 1,164
505000	240	12	3	79	2	8	0	34	1	5	2	0	3	0	0	0	0	0	0	0	2	23	0	1	17	406	1	2	2	0.9976	841 / 843
1010000	428	1	3	111	38	0	0	169	6	1	0	0	6	0	0	0	0	0	0	0	0	1	0	2	0	22	0	0	285	0.7344	788 / 1,073
Total	43510	394	198	19601	5925	1307	0	17136	765	1956	497	0	786	0	0	0	0	0	260	9	273	3367	0	298	845	20045	78	18	744	0.5189	61,242 / 118,012

Fig. 17. Applied Resources Problems - Deep Learning confusion matrix

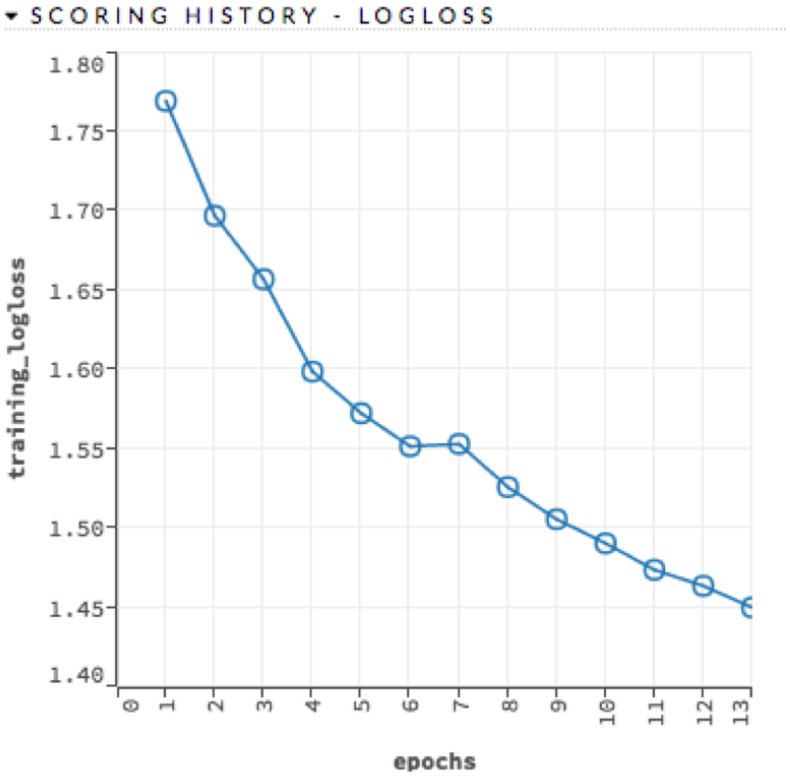


Fig. 18. Applied Resources Problems - DL Logloss evolution

C. Ideal Resources Problems

1) Random Forest: In grid execution on Random Forest, this was the best combination of parameter:

max_depth	min_rows	ntrees	model_ids	logloss
10	3.0	9	trecid_3_grid_model_2	1.8753826655153987
10	1.0	10	trecid_3_grid_model_0	1.881331380507829
20	3.0	8	trecid_3_grid_model_3	2.604809691434048
20	1.0	8	trecid_3_grid_model_1	2.7894270419483727

Fig. 19. Ideal Resources Problems - Grid Results

For "trecid_3_grid_model_2 models", this is the summary:

▾ OUTPUT - CROSS-VALIDATION METRICS SUMMARY

	mean	sd	cv_1_valid	cv_2_valid	cv_3_valid	cv_4_valid	cv_5_valid	cv_6_valid	cv_7_valid	cv_8_valid	cv_9_valid	cv_10_valid
accuracy	0.39763507	0.003975299	0.401092	0.39809737	0.40242046	0.3927242	0.39523277	0.39210504	0.39737955	0.41097644	0.39205617	0.39426675
err	0.60236496	0.003975299	0.598908	0.60190266	0.59757954	0.6072758	0.6047672	0.60789496	0.6026205	0.58902353	0.60794383	0.6057332
err_count	7066.5	55.327435	7130.0	7023.0	7061.0	7178.0	7028.0	7130.0	6991.0	6901.0	7102.0	7121.0
logloss	1.8753947	0.007029275	1.8744	1.8580806	1.8743994	1.8776791	1.8744758	1.8725222	1.894658	1.8618068	1.8839362	1.8819888
max_per_class_error	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
mean_per_class_accuracy	0.1123295	0.0022024254	0.11407437	0.109176226	0.111106955	0.116028465	0.11808718	0.109971814	0.11211666	0.113227755	0.1125819	0.10692363
mean_per_class_error	0.8876705	0.0022024254	0.88592565	0.8908238	0.88889307	0.8839715	0.8819128	0.8900282	0.88788337	0.8867722	0.8874181	0.89307636
mse	0.6160192	0.0015395271	0.61514837	0.6165384	0.61421764	0.6173651	0.61369145	0.6169801	0.6167403	0.61197555	0.61999196	0.61754304
r2	0.9936414	4.163166E-5	0.99360365	0.99369764	0.99369276	0.9935881	0.99374896	0.99360466	0.9936028	0.9935642	0.99369925	0.9936116
rmse	0.7848677	9.80886E-4	0.784314	0.7851996	0.7837204	0.78572583	0.7833846	0.7854808	0.78532815	0.7822887	0.7873957	0.7858391

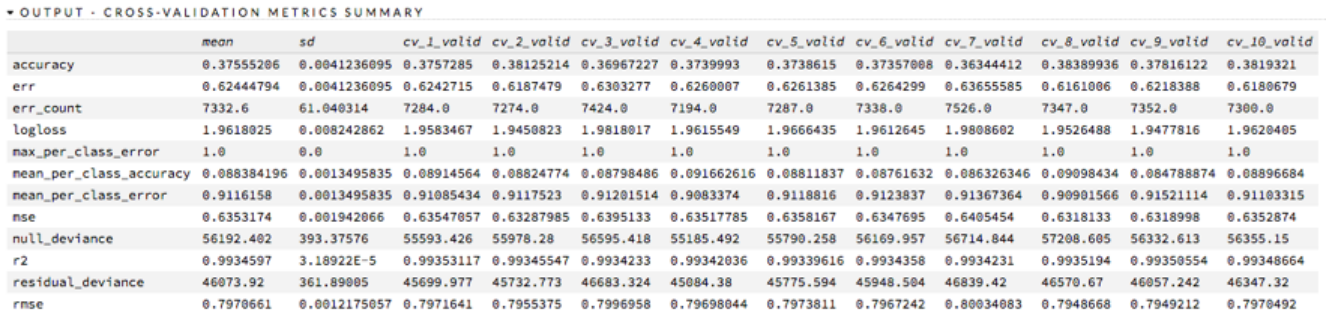
Fig. 20. Ideal Resources Problems - RF summary

▾ CROSS VALIDATION METRICS - CONFUSION MATRIX ROW LABELS: ACTUAL CLASS; COLUMN LABELS: PREDICTED CLASS

	101000	102000	103000	104000	105000	106000	107000	201000	202000	203000	204000	205000	301000	302000	303000	304000	305000	306000	401000	402000	403000	404000	405000	501000	502000	503000	504000	505000	1010000	Error	Rate
101000	19627	6	4	1375	138	19	1	4394	15	68	31	0	8	2	0	2	0	1	10	0	8	52	0	21	22	3197	11	7	31	0.3244	9,423 / 29,050
102000	872	62	0	93	4	0	0	158	0	7	0	0	0	0	0	0	0	0	0	0	0	8	0	1	4	480	0	1	0	0.9633	1,628 / 1,690
103000	844	5	10	196	3	0	0	331	0	3	6	0	2	0	1	0	1	0	5	0	1	23	0	2	10	599	0	7	11	0.9951	2,050 / 2,060
104000	6986	15	4	5730	81	11	3	1234	0	44	44	1	17	0	0	0	0	0	13	1	15	99	0	3	25	1983	3	8	61	0.6502	10,651 / 16,381
105000	2024	3	1	333	271	2	0	3489	1	12	3	0	20	0	0	2	0	0	5	1	2	39	0	4	0	396	0	1	15	0.9591	6,353 / 6,624
106000	2170	3	0	407	1	43	0	67	0	17	0	0	0	0	0	0	0	0	3	0	2	19	0	0	1	317	0	1	0	0.9859	3,008 / 3,051
107000	98	0	0	25	0	1	2	20	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	23	0	0	1	0.9884	170 / 172
201000	2210	1	0	48	224	0	0	10981	12	6	1	0	7	0	0	0	0	0	2	0	0	3	0	17	0	338	0	5	11	0.2081	2,885 / 13,866
202000	397	0	0	11	17	0	0	2369	23	0	1	0	3	0	0	0	0	0	0	0	0	4	0	2	1	65	0	0	3	0.9921	2,873 / 2,896
203000	1987	2	1	277	5	6	0	125	1	256	3	0	0	0	0	0	0	0	2	1	9	43	0	1	14	720	0	0	1	0.9259	3,198 / 3,454
204000	271	0	1	327	13	0	0	159	1	5	30	0	0	0	0	0	0	0	1	1	0	20	0	1	7	263	0	1	1	0.9728	1,072 / 1,102
205000	26	0	0	4	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	38	0	0	0	1.0	100 / 100
301000	301	2	1	13	30	0	0	959	0	0	0	0	63	0	0	0	0	0	6	0	1	5	0	4	2	98	1	1	6	0.9578	1,430 / 1,493
302000	16	0	0	3	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	1	0	1.0	39 / 39
303000	15	0	0	4	3	0	0	8	1	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	3	25	0	0	0	1.0	61 / 61
304000	22	0	0	4	0	0	0	9	0	1	0	0	1	0	0	0	2	0	0	0	0	0	0	0	0	27	0	0	0	0.9697	64 / 66
305000	8	0	0	4	0	0	0	8	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	28	0	1	0	1.0	51 / 51
306000	7	0	0	8	0	0	0	12	0	0	1	0	0	0	0	0	0	1	0	0	0	3	0	0	2	21	0	0	0	0.9818	54 / 55
401000	1103	1	4	180	1	3	0	49	2	16	2	0	4	0	0	0	0	0	28	2	0	21	0	0	2	433	0	0	0	0.9849	1,823 / 1,851
402000	145	0	1	18	1	0	0	11	0	0	0	0	0	0	0	0	0	1	2	4	0	15	0	0	1	119	0	1	0	0.9675	315 / 319
403000	157	0	0	128	2	0	0	70	0	3	0	0	1	0	0	0	0	0	0	0	71	14	0	0	0	78	0	1	0	0.8648	454 / 525
404000	3094	10	4	634	9	6	0	51	0	43	12	0	0	1	0	1	1	1	3	2	12	408	0	0	29	1611	0	0	0	0.9312	5,524 / 5,932
405000	14	0	0	0	2	0	0	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.0	61 / 61
501000	151	1	0	3	13	0	0	62	0	0	1	0	2	0	0	0	0	0	0	0	0	3	0	144	0	92	0	0	0	0.6949	328 / 472
502000	1626	5	3	151	9	4	0	62	0	41	5	0	4	0	0	0	0	0	3	0	1	50	0	0	98	1149	0	0	0	0.9695	3,113 / 3,211
503000	8800	18	4	827	41	8	0	1030	2	46	13	0	6	0	1	1	2	0	3	2	2	81	0	25	48	8680	8	16	9	0.5588	10,993 / 19,673
504000	597	1	1	72	0	2	0	41	0	5	0	0	1	0	0	0	0	0	1	0	0	3	0	0	4	407	1	1	0	0.9991	1,136 / 1,137
505000	298	1	2	53	0	0	0	51	0	1	0	0	1	0	0	0	0	0	0	1	2	6	0	1	9	409	0	23	0	0.9732	835 / 858
1010000	224	0	2	16	8	0	0	703	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	1	0	11	0	0	91	0.9145	973 / 1,064
Total	54090	136	43	10944	876	105	6	26529	58	574	153	1	149	3	2	8	4	4	87	15	126	925	0	228	283	21624	24	76	241	0.6024	70,665 / 117,314

Fig. 21. Ideal Resources Problems - RF confusion matrix

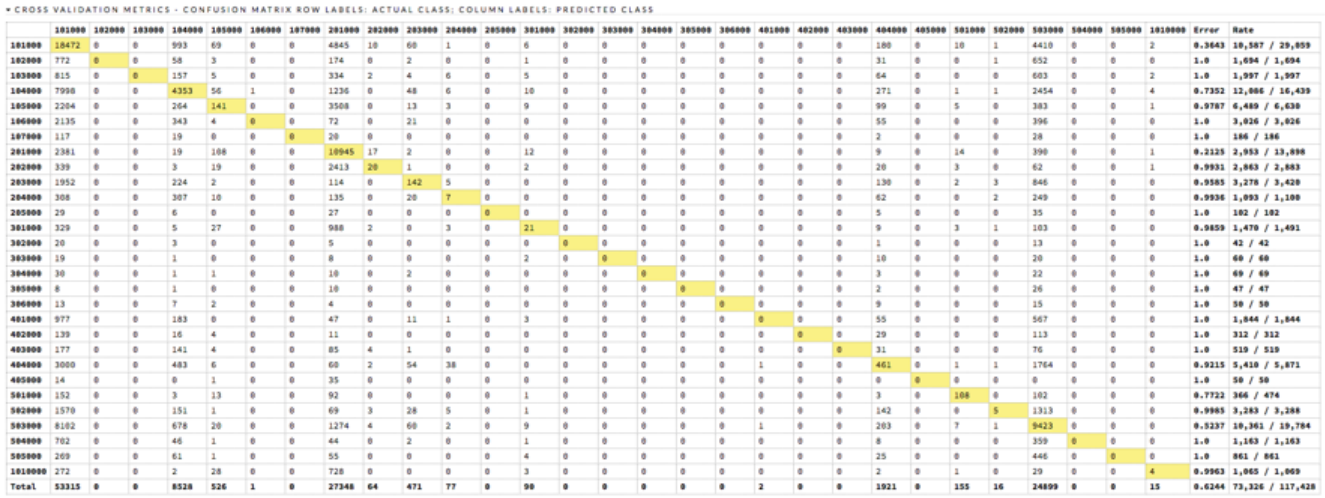
2) Generalized Linear Models: This is the summary for GLM execution



OUTPUT - CROSS-VALIDATION METRICS SUMMARY

	mean	sd	cv_1_valid	cv_2_valid	cv_3_valid	cv_4_valid	cv_5_valid	cv_6_valid	cv_7_valid	cv_8_valid	cv_9_valid	cv_10_valid
accuracy	0.37555206	0.0041236095	0.3757285	0.38125214	0.36967227	0.3739993	0.3738615	0.37357008	0.36344412	0.38389936	0.37816122	0.3819321
err	0.62444794	0.0041236095	0.6242715	0.6187479	0.6303277	0.6260007	0.6261385	0.6264299	0.63655585	0.6161086	0.6218388	0.6180679
err_count	7332.6	61.040314	7284.0	7274.0	7424.0	7194.0	7287.0	7338.0	7526.0	7347.0	7352.0	7300.0
logloss	1.9618025	0.008242862	1.9583467	1.9450823	1.9818017	1.9615549	1.9666435	1.9612645	1.9808602	1.9526488	1.9477816	1.9620405
max_per_class_error	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
mean_per_class_accuracy	0.088384196	0.0013495835	0.08914564	0.08824774	0.08798486	0.091662616	0.08811837	0.08761632	0.086326346	0.09098434	0.084788874	0.08896684
mean_per_class_error	0.9116158	0.0013495835	0.91085434	0.9117523	0.91201514	0.9083374	0.9118816	0.9123837	0.91367364	0.90901566	0.91521114	0.91103315
mse	0.6353174	0.001942866	0.63547857	0.63287985	0.6395133	0.63517785	0.6358167	0.6347695	0.6405454	0.6318133	0.6318998	0.6352874
null_deviance	56192.402	393.37576	55593.426	55978.28	56595.418	55185.492	55790.258	56169.957	56714.844	57208.605	56332.613	56355.15
r2	0.9934597	3.18922E-5	0.99353117	0.99345547	0.9934233	0.99342036	0.99339616	0.9934358	0.9934231	0.9935194	0.99350554	0.99348664
residual_deviance	46073.92	361.89085	45699.977	45732.773	46683.324	45084.38	45775.594	45948.504	46839.42	46570.67	46057.242	46347.32
rmse	0.7970661	0.0012175057	0.7971641	0.7955375	0.7996958	0.79698044	0.7973811	0.7967242	0.80034083	0.7948668	0.7949212	0.7970492

Fig. 22. Ideal Resources Problems - GLM summary

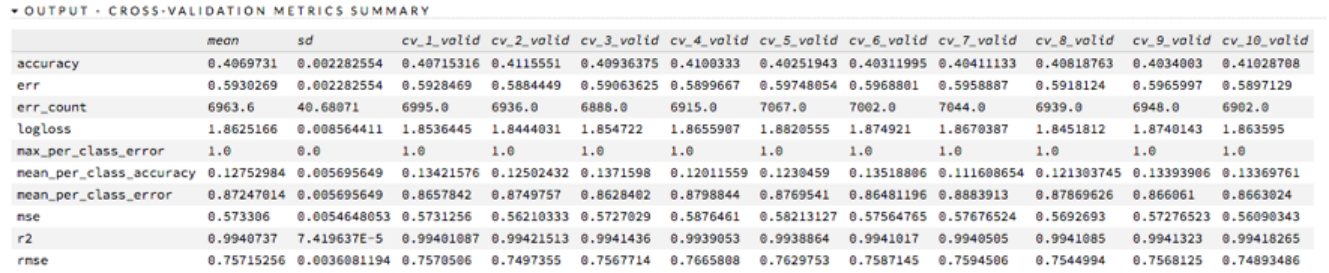


CROSS VALIDATION METRICS - CONFUSION MATRIX ROW LABELS: ACTUAL CLASS; COLUMN LABELS: PREDICTED CLASS

	101000	102000	103000	104000	105000	106000	107000	201000	202000	203000	204000	205000	301000	302000	303000	304000	305000	306000	401000	402000	403000	404000	405000	501000	502000	503000	504000	505000	1010000	Error	Rate
101000	18472	0	0	993	69	0	0	4845	10	60	1	0	6	0	0	0	0	0	0	0	0	180	0	10	1	4410	0	0	2	0.3643	10,587 / 29,059
102000	772	0	0	58	3	0	0	174	0	2	0	0	1	0	0	0	0	0	0	0	0	31	0	0	1	652	0	0	0	1.0	1,694 / 1,694
103000	815	0	0	157	5	0	0	334	2	4	6	0	5	0	0	0	0	0	0	0	0	64	0	0	0	603	0	0	2	1.0	1,997 / 1,997
104000	7998	0	0	4353	56	1	0	1236	0	48	6	0	10	0	0	0	0	0	0	0	0	271	0	1	1	2454	0	0	4	0.7352	12,086 / 16,439
105000	2204	0	0	264	141	0	0	3508	0	13	3	0	9	0	0	0	0	0	0	0	0	99	0	5	0	383	0	0	1	0.9787	6,489 / 6,630
106000	2135	0	0	343	4	0	0	72	0	21	0	0	0	0	0	0	0	0	0	0	0	55	0	0	0	396	0	0	0	1.0	3,026 / 3,026
107000	117	0	0	19	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	28	0	0	0	1.0	186 / 186
201000	2381	0	0	19	108	0	0	10945	17	2	0	0	12	0	0	0	0	0	0	0	0	9	0	14	0	390	0	0	1	0.2125	2,953 / 13,898
202000	339	0	0	3	19	0	0	2413	20	1	0	0	2	0	0	0	0	0	0	0	0	20	0	3	0	62	0	0	1	0.9931	2,863 / 2,883
203000	1952	0	0	224	2	0	0	114	0	142	5	0	0	0	0	0	0	0	0	0	0	130	0	2	3	846	0	0	0	0.9585	3,278 / 3,420
204000	308	0	0	307	10	0	0	135	0	20	7	0	0	0	0	0	0	0	0	0	0	62	0	0	2	249	0	0	0	0.9936	1,093 / 1,100
205000	29	0	0	6	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	35	0	0	0	1.0	102 / 102
301000	329	0	0	5	27	0	0	988	2	0	3	0	21	0	0	0	0	0	0	0	0	9	0	3	1	103	0	0	0	0.9859	1,470 / 1,491
302000	20	0	0	3	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	13	0	0	0	1.0	42 / 42
303000	19	0	0	1	0	0	0	8	0	0	0	0	2	0	0	0	0	0	0	0	0	10	0	0	0	20	0	0	0	1.0	60 / 60
304000	30	0	0	1	1	0	0	10	0	2	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	22	0	0	0	1.0	69 / 69
305000	8	0	0	1	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	26	0	0	0	1.0	47 / 47
306000	13	0	0	7	2	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	15	0	0	0	1.0	50 / 50
401000	977	0	0	183	0	0	0	47	0	11	1	0	3	0	0	0	0	0	0	0	0	55	0	0	0	567	0	0	0	1.0	1,844 / 1,844
402000	139	0	0	16	4	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	29	0	0	0	113	0	0	0	1.0	312 / 312
403000	177	0	0	141	4	0	0	85	4	1	0	0	0	0	0	0	0	0	0	0	0	31	0	0	0	76	0	0	0	1.0	519 / 519
404000	3000	0	0	483	6	0	0	60	2	54	38	0	0	0	0	0	0	0	0	1	0	0	461	0	1	1764	0	0	0	0.9215	5,410 / 5,871
405000	14	0	0	0	1	0	0	35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.0	50 / 50
501000	152	0	0	3	13	0	0	92	0	0	0	0	1	0	0	0	0	0	0	0	0	3	0	108	0	102	0	0	0	0.7722	366 / 474
502000	1570	0	0	151	1	0	0	69	3	28	5	0	1	0	0	0	0	0	0	0	0	142	0	0	5	1313	0	0	0	0.9985	3,283 / 3,288
503000	8102	0	0	678	20	0	0	1274	4	60	2	0	9	0	0	0	0	0	0	1	0	203	0	7	1	9423	0	0	0	0.5237	10,361 / 19,784
504000	702	0	0	46	1	0	0	44	0	2	0	0	1	0	0	0	0	0	0	0	0	8	0	0	0	359	0	0	0	1.0	1,163 / 1,163
505000	269	0	0	61	1	0	0	55	0	0	0	0	4	0	0	0	0	0	0	0	0	25	0	0	0	446	0	0	0	1.0	861 / 861
1010000	272	0	0	2	28	0	0	728	0	0	0	0	3	0	0	0	0	0	0	0	0	2	0	1	0	29	0	0	4	0.0963	1,065 / 1,069
Total	53315	0	0	8528	526	1	0	27348	64	471	77	0	98	0	0	0	0	0	2	0	0	1921	0	155	16	24899	0	0	15	0.6244	73,326 / 117,428

Fig. 23. Ideal Resources Problems - GLM confusion matrix

3) Deep Learning - Neural Nets: This is the summary for Deep Learning execution



OUTPUT - CROSS-VALIDATION METRICS SUMMARY

	mean	sd	cv_1_valid	cv_2_valid	cv_3_valid	cv_4_valid	cv_5_valid	cv_6_valid	cv_7_valid	cv_8_valid	cv_9_valid	cv_10_valid
accuracy	0.4069731	0.002282554	0.40715316	0.4115551	0.40936375	0.4100333	0.40251943	0.40311995	0.40411133	0.40818763	0.4034003	0.41028708
err	0.5930269	0.002282554	0.5928469	0.5884449	0.59063625	0.5899667	0.59748054	0.5968801	0.5958887	0.5918124	0.5965997	0.5897129
err_count	6963.6	40.68071	6995.0	6936.0	6888.0	6915.0	7067.0	7002.0	7044.0	6939.0	6948.0	6902.0
logloss	1.8625166	0.008564411	1.8536445	1.8444031	1.854722	1.8655907	1.8820555	1.874921	1.8670387	1.8451812	1.8740143	1.863595
max_per_class_error	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
mean_per_class_accuracy	0.12752984	0.005695649	0.13421576	0.12502432	0.1371598	0.12011559	0.1230459	0.13518806	0.111608654	0.121303745	0.13393906	0.13369761
mean_per_class_error	0.87247014	0.005695649	0.8657842	0.8749757	0.8628402	0.8798844	0.8769541	0.86481196	0.8883913	0.87869626	0.866061	0.8663024
mse	0.573306	0.0054648053	0.5731256	0.56210333	0.5727029	0.5876461	0.58213127	0.57564765	0.57676524	0.5692693	0.57276523	0.56090343
r2	0.9940737	7.419637E-5	0.99401087	0.99421513	0.9941436	0.9939053	0.9938864	0.9941017	0.9940505	0.9941085	0.9941323	0.99418265
rmse	0.75715256	0.0036081194	0.7570506	0.7497355	0.7567714	0.7665808	0.7629753	0.7587145	0.7594506	0.7544994	0.7568125	0.74893486

Fig. 24. Ideal Resources Problems - Deep Learning summary

▾ CROSS VALIDATION METRICS - CONFUSION MATRIX ROW LABELS: ACTUAL CLASS; COLUMN LABELS: PREDICTED CLASS

	101000	102000	103000	104000	105000	106000	107000	201000	202000	203000	204000	205000	301000	302000	303000	304000	305000	306000	401000	402000	403000	404000	405000	501000	502000	503000	504000	505000	1010000	Error	Rate
101000	19708	16	4	1569	535	29	0	3271	54	228	35	0	31	0	0	0	0	0	33	0	27	339	0	22	29	2958	4	1	82	0.3198	9,267 / 28,975
102000	811	82	0	123	33	1	0	128	1	8	5	0	0	0	0	0	0	0	2	1	1	33	0	0	9	456	0	0	0	0.9516	1,612 / 1,694
103000	859	6	8	227	21	6	0	215	1	19	15	0	8	0	0	0	0	0	22	0	3	57	0	1	19	553	0	4	19	0.9961	2,055 / 2,063
104000	6698	24	15	6055	317	31	0	703	6	88	85	0	23	0	0	0	0	0	24	0	67	396	0	3	35	1796	2	3	87	0.6321	10,403 / 16,458
105000	1905	7	2	405	1373	6	0	2370	11	25	10	0	23	0	0	0	0	0	3	0	4	125	0	5	13	346	1	0	50	0.7946	5,311 / 6,684
106000	2064	3	1	414	4	61	0	49	0	24	1	0	3	0	0	0	0	0	8	0	12	73	0	0	5	313	2	0	2	0.9799	2,978 / 3,039
107000	181	0	0	30	2	1	0	15	0	5	0	0	0	0	0	0	0	0	0	0	1	5	0	0	0	19	0	0	1	1.0	180 / 180
201000	2493	4	3	128	1065	2	0	9562	124	13	1	0	45	0	0	0	0	0	3	0	2	18	0	12	1	446	0	0	59	0.3161	4,419 / 13,981
202000	462	0	2	14	233	0	0	1940	116	1	0	0	9	0	0	0	0	0	0	1	2	10	0	6	4	83	0	0	16	0.9600	2,783 / 2,899
203000	1634	3	0	249	7	19	0	112	3	531	8	0	2	0	0	0	0	0	23	0	20	184	0	0	31	641	1	0	1	0.8469	2,938 / 3,469
204000	263	1	3	275	41	5	0	129	1	24	82	0	0	0	0	0	0	0	5	0	2	52	0	0	15	201	0	0	2	0.9255	1,019 / 1,101
205000	25	0	1	7	1	0	0	28	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	4	32	0	0	0	1.0	104 / 104
301000	347	1	0	40	104	0	0	707	4	1	1	0	106	0	0	0	0	0	4	0	1	16	0	2	0	119	0	0	15	0.9278	1,362 / 1,468
302000	19	0	0	2	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	1.0	36 / 36
303000	18	0	0	2	2	0	0	6	0	1	0	0	1	0	0	0	0	0	0	0	0	5	0	0	4	14	0	0	0	1.0	53 / 53
304000	22	0	0	5	1	0	0	7	0	2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	29	0	0	0	1.0	68 / 68
305000	16	0	1	7	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	19	0	0	0	1.0	52 / 52
306000	11	0	0	5	0	0	0	5	0	1	1	0	0	0	0	0	0	0	0	0	0	7	0	0	6	13	0	0	0	1.0	49 / 49
401000	968	1	6	199	4	2	0	38	1	51	3	0	8	0	0	0	0	0	65	0	3	74	0	1	20	414	0	0	0	0.9650	1,793 / 1,858
402000	128	0	0	23	4	1	0	6	0	0	1	0	0	0	0	0	0	0	2	3	1	38	0	0	2	104	0	0	0	0.9904	310 / 313
403000	120	0	0	130	0	1	0	42	5	6	3	0	0	0	0	0	0	0	0	0	125	30	0	0	5	68	0	1	0	0.7668	411 / 536
404000	2471	18	4	741	19	12	0	43	0	115	23	0	0	0	0	0	0	0	22	0	29	904	0	1	78	1363	1	1	1	0.8454	4,942 / 5,846
405000	19	0	0	0	3	0	0	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1.0	60 / 60
501000	154	0	0	7	17	0	0	54	0	1	0	0	0	0	0	0	0	0	0	0	2	5	0	131	1	99	0	0	4	0.7242	344 / 475
502000	1457	18	1	145	33	6	0	47	0	92	6	0	2	0	0	0	0	0	22	2	5	135	0	1	173	1125	1	1	1	0.9471	3,100 / 3,273
503000	8507	78	3	800	95	15	0	833	2	164	16	0	15	0	0	0	0	0	30	4	16	336	0	16	102	8545	7	6	14	0.5641	11,059 / 19,604
504000	639	3	0	75	5	0	0	44	0	9	1	0	0	0	0	0	0	0	0	0	0	10	0	1	1	352	2	1	0	0.9963	1,141 / 1,143
505000	286	3	2	71	3	2	0	40	0	2	0	0	4	0	0	0	0	0	0	0	2	27	0	0	17	486	1	6	0	0.9931	866 / 872
1010000	384	1	5	106	96	0	0	292	3	0	0	0	11	0	0	0	0	0	0	0	1	3	0	3	0	16	0	0	150	0.8599	921 / 1,071
Total	52589	269	61	11854	4018	200	0	20733	332	1411	297	0	293	0	0	0	0	0	268	11	326	2890	0	205	574	20542	22	24	505	0.5930	69,636 / 117,424

Fig. 25. Ideal Resources Problems - Deep Learning confusion matrix

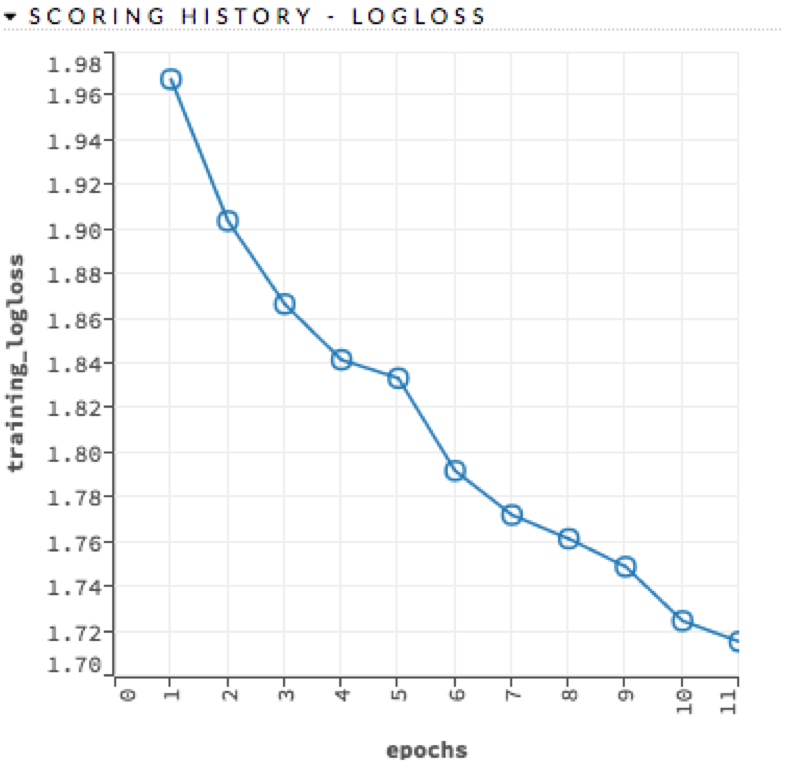


Fig. 26. Ideal Resources Problems - DL Logloss evolution

D. Abuse problems

1) Random Forest: In grid execution on Random Forest, this was the best combination of parameter:

max_depth	min_rows	ntrees	model_ids	logloss
40	3.0	51	tvalin_maltrato_grid_2_model_15	0.04463371858532171
40	1.0	46	tvalin_maltrato_grid_2_model_13	0.04599511129235313
40	3.0	47	tvalin_maltrato_grid_2_model_9	0.04636275050907876
40	1.0	48	tvalin_maltrato_grid_2_model_7	0.046440100259489414
40	1.0	44	tvalin_maltrato_grid_2_model_1	0.04654999968936572
40	5.0	47	tvalin_maltrato_grid_2_model_5	0.04667900221579643
40	5.0	45	tvalin_maltrato_grid_2_model_11	0.04679377561405307
40	5.0	56	tvalin_maltrato_grid_2_model_17	0.0472961990961147
40	3.0	46	tvalin_maltrato_grid_2_model_3	0.04744548223349003
20	3.0	64	tvalin_maltrato_grid_2_model_14	0.06462231774040969
20	1.0	67	tvalin_maltrato_grid_2_model_12	0.06462705436126337
20	1.0	49	tvalin_maltrato_grid_2_model_6	0.0658182554180864
20	5.0	50	tvalin_maltrato_grid_2_model_4	0.06591471539094428
20	3.0	57	tvalin_maltrato_grid_2_model_8	0.06657333868892994
20	1.0	50	tvalin_maltrato_grid_2_model_0	0.06661047962804245
20	5.0	61	tvalin_maltrato_grid_2_model_16	0.06673884149010913
20	5.0	62	tvalin_maltrato_grid_2_model_10	0.06718412286238035
20	3.0	43	tvalin_maltrato_grid_2_model_2	0.067865282691744

Fig. 27. Abuse Problems - Grid Results

For ”tvalin_maltrato_grid_2_model_15 models”, this is the summary:

OUTPUT - CROSS-VALIDATION METRICS SUMMARY

	mean	sd	cv_1_valid	cv_2_valid	cv_3_valid	cv_4_valid	cv_5_valid	cv_6_valid	cv_7_valid	cv_8_valid	cv_9_valid	cv_10_valid
accuracy	0.9981803	3.7251785E-4	0.99832636	0.9987087	0.9986006	0.9970195	0.998133	0.99846023	0.9980769	0.99860674	0.99847364	0.99748565
auc	0.99437803	0.0020108677	0.9960638	0.9987793	0.9972555	0.99129164	0.993189	0.9924913	0.995185	0.9955055	0.99538153	0.9886378
err	0.0018196553	3.7251785E-4	0.0016736402	0.0012993149	0.001399417	0.0029804483	0.0018669778	0.0015397371	0.0019230769	0.0013932427	0.001526359	0.0025943397
err_count	15.4	3.1016126	14.0	11.0	12.0	25.0	16.0	13.0	16.0	12.0	13.0	22.0
f0point5	0.98994535	0.0017249048	0.98617756	0.9927983	0.99353796	0.9871134	0.9891984	0.99130905	0.9897519	0.99071205	0.9919436	0.986911
f1	0.9798561	0.0039210944	0.9814815	0.9859515	0.984	0.96839446	0.9775281	0.98250335	0.97866666	0.9846154	0.9837703	0.97164947
f2	0.970083	0.0065274225	0.9768299	0.9791984	0.9746434	0.9503722	0.9661299	0.9738527	0.967827	0.9785933	0.97573054	0.9568528
lift_top_group	21.832848	0.5822432	21.955381	21.378788	22.506561	20.558823	23.608816	22.395226	21.780106	21.860407	20.977833	21.306532
logloss	0.044653766	0.002059558	0.04530701	0.044091403	0.042884745	0.049135234	0.04063228	0.04463346	0.045202136	0.039518118	0.04683806	0.048295215
max_per_class_error	0.036439057	0.008244672	0.026246719	0.025252525	0.031496063	0.06127451	0.041322313	0.03183824	0.039267015	0.02538071	0.02955665	0.05276382
mcc	0.9790834	0.00401701	0.980639	0.98534197	0.9834061	0.9673644	0.9767601	0.981816	0.9778469	0.98394257	0.9830717	0.9706451
mean_per_class_accuracy	0.98170614	0.0040994035	0.98662615	0.9873118	0.984252	0.96936274	0.9792779	0.9840229	0.9803035	0.987188	0.98516005	0.9735562
mean_per_class_error	0.018293882	0.0040994035	0.01337386	0.0126882205	0.015748031	0.030637255	0.02072208	0.015977109	0.019696496	0.012812025	0.01483997	0.026443776
nse	0.01355051	6.4846716E-4	0.014222522	0.01437307	0.013139136	0.014057338	0.011812746	0.013668495	0.013751962	0.012003458	0.014708499	0.013767876
precision	0.9968076	0.0020230378	0.98933333	0.997416	1.0	1.0	0.9971347	0.9972678	0.9972826	0.9948186	0.99746835	0.9973545
r2	0.69048977	0.010926406	0.6728379	0.6776428	0.69053316	0.6962216	0.7087799	0.6795836	0.6860671	0.72502065	0.67600286	0.6922084
recall	0.96356094	0.008244672	0.9737533	0.9747475	0.96850395	0.9387255	0.9586777	0.96816975	0.960733	0.97461927	0.97044337	0.9472362
rmse	0.11633787	0.0028293005	0.11925822	0.11988774	0.11462607	0.118563645	0.108686455	0.11691234	0.117268756	0.109560296	0.1212786	0.11733659
specificity	0.9998513	9.5138996E-5	0.999499	0.9998761	1.0	1.0	0.99987817	0.999876	0.999874	0.99975663	0.99987674	0.99987626

Fig. 28. Abuse Problems - RF summary

	0	1	Error	Rate
0	20213	7	0.0003	(7.0/20220.0)
1	51	864	0.0557	(51.0/915.0)
Total	20264	871	0.0027	(58.0/21135.0)

Fig. 29. Abuse Problems - RF confusion matrix

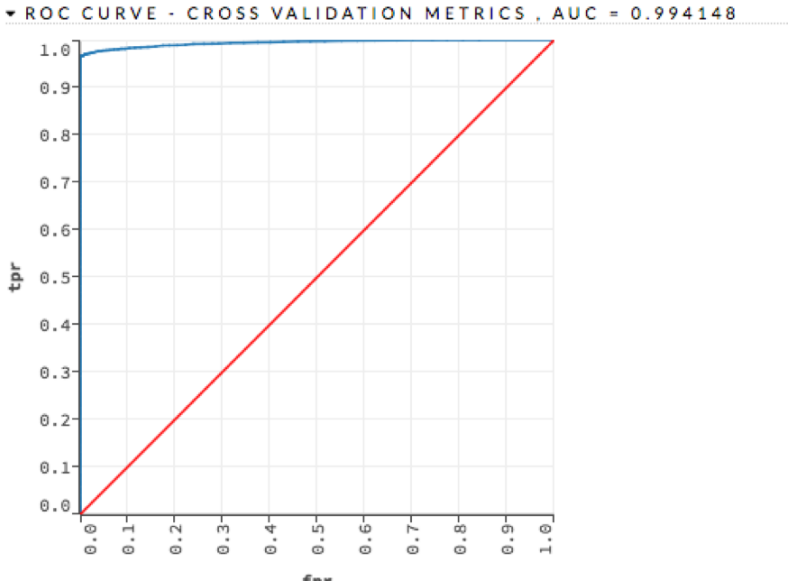


Fig. 30. Abuse Problems - RF ROC Curve

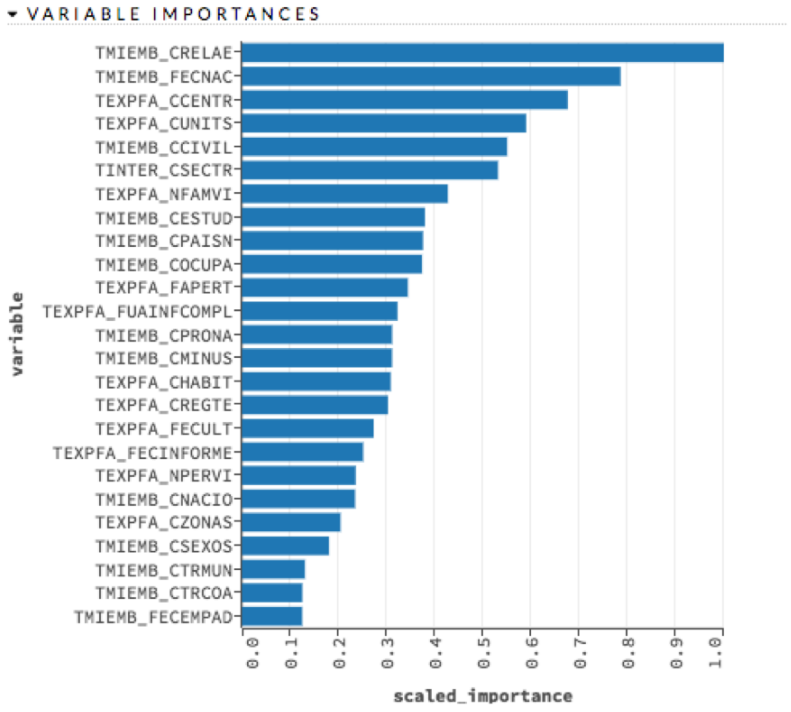


Fig. 31. Abuse Problems - RF Variables importances

E. Resources matching problems

1) Random Forest: In grid execution on Random Forest, this was the best combination of parameter:

max_depth	min_rows	ntrees	model_ids	logloss
40	1.0	20	trecid_coin_grid_model_1	0.06813358472611751
40	1.0	46	trecid_coin_grid_model_7	0.0694517341734561
40	3.0	20	trecid_coin_grid_model_3	0.07221660819841366
40	5.0	20	trecid_coin_grid_model_5	0.0767496812294146
40	5.0	49	trecid_coin_grid_model_11	0.08137917916471142
40	3.0	30	trecid_coin_grid_model_9	0.08636001188158178
20	1.0	50	trecid_coin_grid_model_6	0.08648604264172752
20	3.0	38	trecid_coin_grid_model_8	0.09266601608546655
20	5.0	20	trecid_coin_grid_model_4	0.09337935173178252
20	3.0	20	trecid_coin_grid_model_2	0.09505348362165038
20	1.0	19	trecid_coin_grid_model_0	0.09671126238510172
20	5.0	18	trecid_coin_grid_model_10	0.11481076418456221

Fig. 32. Resources Matching Problems - Grid Results

For ”trecid_coin_grid_model_1 models”, this is the summary:

▾ OUTPUT · CROSS-VALIDATION METRICS SUMMARY

	mean	sd	cv_1_valid	cv_2_valid	cv_3_valid	cv_4_valid	cv_5_valid
accuracy	0.9961116	3.9390556E-4	0.9966382	0.99686205	0.9953982	0.99568254	0.9959768
auc	0.99445474	0.001038282	0.9933721	0.9965781	0.99491477	0.9923329	0.9950758
err	0.0038884473	3.9390556E-4	0.0033618403	0.0031379515	0.00460177	0.004317483	0.0040231925
err_count	65.8	6.68431	57.0	53.0	78.0	73.0	68.0
f0point5	0.9967604	3.3641767E-4	0.9971588	0.9974367	0.99614197	0.9964116	0.99665296
f1	0.9978911	2.1240472E-4	0.9981744	0.9983004	0.9975137	0.99765825	0.997809
f2	0.9990245	9.114449E-5	0.99919206	0.9991655	0.9988892	0.9989079	0.99896777
lift_top_group	1.0865448	0.0017873357	1.0871722	1.084848	1.083067	1.0870515	1.0905849
logloss	0.06812983	0.0031736486	0.07067722	0.059213854	0.069435805	0.07033242	0.07098983
max_per_class_error	0.046230637	0.005245895	0.040145986	0.03709311	0.057692308	0.05096012	0.04526167
mcc	0.97325414	0.0029075677	0.9771862	0.9780651	0.9671195	0.970385	0.973515
mean_per_class_accuracy	0.97677547	0.002623342	0.97986287	0.981325	0.971058	0.97439134	0.97724
mean_per_class_error	0.023224557	0.002623342	0.020137157	0.018675016	0.028942	0.025608644	0.022759967
mse	0.020899935	8.177957E-4	0.020970097	0.018663475	0.021843791	0.021658089	0.021364223
precision	0.9960081	4.191732E-4	0.9964829	0.9968618	0.9952296	0.9955823	0.9958837
r2	0.715527	0.011661057	0.71766245	0.74112606	0.6915323	0.70600235	0.7213118
recall	0.99978155	3.6608908E-5	0.9998717	0.9997431	0.9998083	0.9997428	0.99974173
rmse	0.14451066	0.0028815076	0.14481056	0.13661434	0.14779645	0.14716688	0.14616506
specificity	0.9537694	0.005245895	0.959854	0.9629069	0.9423077	0.9490399	0.9547383

Fig. 33. Resources Matching Problems - RF summary

	0	1	Error	Rate
0	1632	54	0. 032	(54.0/1686.0)
1	6	19575	0. 0003	(6.0/19581.0)
Total	1638	19629	0.0028	(60.0/21267.0)

Fig. 34. Resources Matching Problems - RF confusion matrix

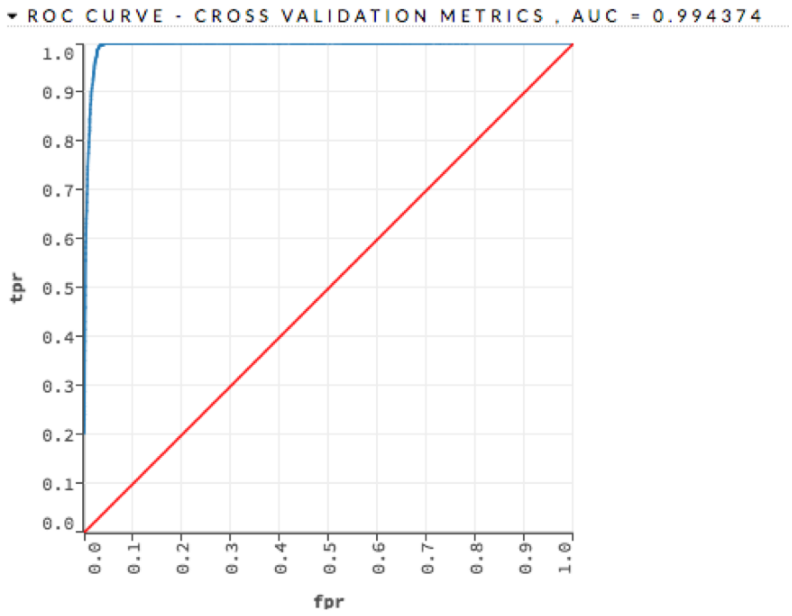


Fig. 35. Resources Matching Problems - RF ROC Curve

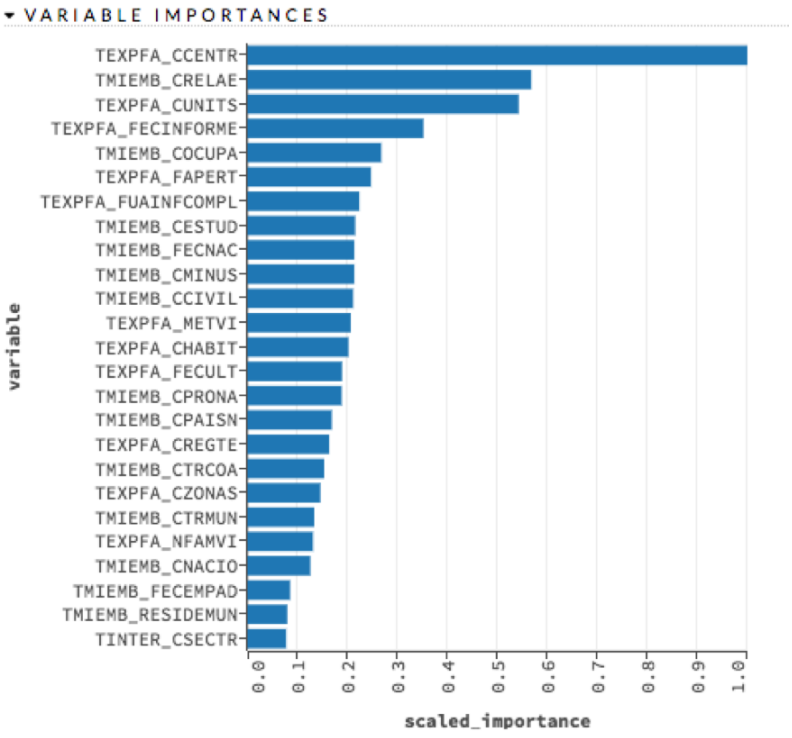


Fig. 36. Resources Matching Problems - RF Variables importances

F. Results Summary

1) Multinomial Problems: For multinomial problems, we can observe the best performance is obtained with Deep Learning model.

2) Binomial Problems: For binomial problems, we can observe that Random Forest model get a really good performance.

VI. Conclusions

Finally, after the development of all steps discussed in previous sections, a first approximation to a Big Data system with Machine Learning in the social services scope has been obtained. This could be very useful for all social workers, in the city of Málaga and, potentially, in all Spain.

Social workers, could have a final application that helps them make better decisions in their attention to new users of the social services and, also, people who already use it. They could provide a better service and anticipate to possible needs and demands of the users.

In this project, the entire cycle of life of the system has been developed. The analysis of the existing data model, the process of extracting, transforming and loading from it, the building of a set of machine learning algorithms and a real time prediction service as well as a web application for its better use and visualization.

Of course, this is just the first stone that can be done. More prediction problems can be added, improve Machine Learning algorithms to get more reliable results, add semantic analyze to process and categorize descriptive fields, etc. The possibilities are endless.

Nevertheless, honestly, I think that the obtained result is more than satisfactory, since, despite the complexity of the data model and its peculiarities, good predictions are obtained with the machine learning models developed, and, also, there is a tool easy to handle, that use these models to make real time predictions.

APPENDIX A
SIUSS TABLES DETAILS

TABLE XIII
TABLE TEXPFA

Name	Description
CPROVI	Código provincia
CCENTR	Código CSS
CUNITS	Código UTS
CODCOA	Código de CC.AA.
FAPERT	Fecha de apertura
NPERVI	N Personas en la Vivienda
NFAMVI	N Familiares en la VIvienda
CREGTE	Código Régimen de Tenencia
DOMICI	Domicilio (anonimizado)
CMUNIC	Código Municipal
CZONAS	Código Zona
DPOBLA	Población
NTELEF	N de Teléfono
CHABIT	Código de N Habitaciones
CTIPVI	Código de tipo de vivienda
FECULT	Fecha Ultima Actualización del Expediente
CEXCOM	Número de Expediente Completo: N Provincia(2) + N CSS (3) + N UTS(2) + N Secuencia (6 Dígitos)
METVI	Metros Cuadrados de la vivienda
PTEBORR	Expediente pendiente de borrado por transpaso a otra UTS
QMEDIO	Ingresos Medios Familiares (Campo calculado)
NTELEF2	Número de teléfono secundario (anonimizado)
NUMDOM	Número de domicilio
ENCRIP	-
SINVIV	-
COMPLETO	-
FECALTA	Fecha de alta
CODPOS	Código postal
COSTEVIV	Coste Anual de la Vivienda
FUAINFCOMPL	Fecha de última actualización informe completado
FECINFORME	Fecha informe
INFORMEMOTIVO	Motivo del informe
INFORMEDERIVADO	Informe derivado a
INFORMECODPROFAL	Código del profesional que emite el informe
CODRESPONSEXP	Código de responsable de expediente
DIAGNSOCIAL	Diagnóstico social
PRONOSTICO	Pronóstico
PROPS	Propuestas a corto, medio y largo plazo
NUMEXPMUN	Número de expediente municipal
OBSERV	Observaciones
SITECON	Situación económica, distribución del presupuesto familiar
INFORMESOLICITADO	Informe solicitado por
HISTFAM	Historia familiar
RELPADRESHIJOS	Relación padre/hijos
VALORAPOYOOTRAS	Valoración del tipo de apoyo a la familia
EQUIPA	Descripción de los equipamientos y recursos del entorno
RELMUF	Relaciones sociales de los miembros adultos de la unidad familiar
ROLES	Reparto y desempeño de roles
ENFERMEDADES	Enfermedades y minusvalías
PERSCUIDMEN	Personas que ejercen de cuidadores de los menores
ESTABEMOC	Estabilidad emocional
RELPAREJA	Relación con la pareja
LIMITACIONES	Limitaciones que no constituyen minusvalías
NORMAS	Normas y pautas
RELHERMANOS	Relación entre hermanos
ABUSOS	Abusos de drogas, alcohol y fármacos
RELFAMESC	Relación familia/escuela
RELNINOS	Relaciones sociales de los niños
ATNNECMED	Atención de las necesidades médicas de lops niños (vacunas y revisiones)
GRADOATN	Grado de atención de necesidades emocionales de los menores
PLANFAM	Planificación Familiar

TABLE XIV
TABLE TMIEMB

Name	Description
CEXCOM	Número de Expediente Completo: N Provincia(2) + N CSS (3) + N UTS(2) + N Secuencia (6 Dígitos)
CINDFA	Código Individual Familiar
CTRMUN	Año de Inicio de Residencia en Municipio (aaaa)
DAPEL1	Apellido (anonimizado)
DAPEL2	Apellido (anonimizado)
DNOMBR	Nombre (anonimizado)
NUMDNI	DNI (anonimizado)
NUMTARJSANIT	Número de tarjeta sanitaria (anonimizado)
TITINFORME	Es titular del último informe elaborado?
PASAPORTE	Pasaporte (anonimizado)
MEDICO	Médico
CTRSAL	Centro de Salud
HORCONS	Horario consulta
DIAGNDISCAP	Diagnóstico de Discapacidad
CURCENTRO	Curso y Centro Escolar
SITHISTLAB	Situación e historia laboral
DIRCTRSAL	Dirección del centro de salud
OBSERV	Observaciones
NTELEF	Número de teléfono
TELCTRSAL	Número de teléfono de centro de salud
NTELEF2	Número de teléfono secundario

TABLE XV
TABLE TINTER

Name	Description
CEXCOM	Número de Expediente Completo: N Provincia(2) + N CSS (3) + N UTS(2) + N Secuencia (6 Dígitos)
NINTER	N de Intervención
CSECTR	Código de sector de referencia
CESTINT	Código del Estado de la Intervención
FECFUA	Fecha de última actualización de la intervención
FECINI	Fecha de inicio de la intervención
FECFIN	Fecha fin de la intervención
COMPLETO	Código de indicador de completado
FECALTA	Fecha de alta
CODRESPONSINT	Código Responsable de Intervención
CARGOSISDEP	Con cargo al sistema nacional de dependencia?
FECSOLVAL	Fecha de solicitud de valoración
FECVAL	Fecha de valoración y PIA teórico
FECACUERDO	Fecha de firma acuerdo con el usuario
FECSALACUERDO	Fecha de salida del acuerdo
FECRESPIA	Fecha de resolución del PIA
CAMBIOS	Con cambios?
OBSER	Observaciones

TABLE XVI
TABLE TUSUIN

Name	Description
CEXCOM	Número de Expediente Completo: N Provincia(2) + N CSS (3) + N UTS(2) + N Secuencia (6 Dígitos)
NINTER	N de Intervención
CINDFA	Código Individual Familiar
FECALTA	Fecha de alta

TABLE XVII
TABLE TVALIN

Name	Description
CEXCOM	Número de Expediente Completo: N Provincia(2) + N CSS (3) + N UTS(2) + N Secuencia (6 Dígitos)
NINTER	N de Intervención
CODCOM	Código de Recurso
FECALTA	Fecha de alta

TABLE XVIII
TABLE TDEMIN

Name	Description
CEXCOM	Número de Expediente Completo: N Provincia(2) + N CSS (3) + N UTS(2) + N Secuencia (6 Dígitos)
NINTER	N de Intervención
CODCOM	Código de Recurso
FECALTA	Fecha de alta

TABLE XIX
TABLE TRECID

Name	Description
CEXCOM	Número de Expediente Completo: N Provincia(2) + N CSS (3) + N UTS(2) + N Secuencia (6 Dígitos)
NINTER	N de Intervención
CODCOM	Código de Recurso Idóneo
FECALTA	Fecha de alta
CODNCOIN	Código de no coincidencia recurso idoneo y recurso aplicado

TABLE XX
TABLE TRECAP

Name	Description
CEXCOM	Número de Expediente Completo: N Provincia(2) + N CSS (3) + N UTS(2) + N Secuencia (6 Dígitos)
NINTER	N de Intervención
CODCOM	Código de Recurso Aplicado
CESTAD	Estado del Recurso
FECALTA	Fecha de alta
DESTINO	Lugar de Destino de los Recursos Aplicados con estado Derivado
FECRES	Fecha de Resolución
CODTIPRES	Código de Tipo de Resolución
CODTIPAYUDA	Código de Tipo de Ayuda
AYUPAGUNICO	Ayuda Pago único (Euros)
NOMAPESPERCEP	Nombre y apellidos del Perceptor
DNIPERCEP	DNI del perceptor
DOMPERCEP	Domicilio del perceptor
LOCPERCEP	Localidad del percetor
TELPERCEP	Teléfono del perceptor
CODPROVPERCEP	Código provincia del perceptor
RELPERCEPUSUS	Relación del Perceptor con los Usuarios
FECINIPRE	Fecha de Inicio de Prestación
CODPERCONC	Código Periodo de Concesión
AYUPERMENS	Ayuda Periódica Mensual (Euros)
FECFINPREST	Fecha fin de la percepción
CODCAUSAFINPREST	Código de causa de fin de prestaciones
CINDFA	Cindfa, del titular del Recurso-Prestación
FECSOLPREST	Fecha de solicitud de la prestación
ENTIDAD	Entidad/ Servicio donde se tramite el Recurso
CODMOTDENEG	Código Motivo de Denegación
OBSERV	Observaciones
FECSALREG	Fecha de Salida de Registro
NUMCTAPERCEP	N de cuenta del Perceptor
CODTIPRESOTRAPREST	Código de tipo de resolución de otra prestación
CODMOTDENEGOTRAPREST	Código de motivo de denegación de otra prestación
CODTIPRESALOJ	Código tipo prestación de alojamiento
CODMOTDENEGALOJ	Código motivo denegación alojamiento
OBSERVALOJ	Observaciones alojamiento
CODFORMAINGRALOJ	Código forma ingreso de alojamiento

TABLE XXI
TABLE TGESINT

Name	Description
CEXCOM	Número de Expediente Completo: N Provincia(2) + N CSS (3) + N UTS(2) + N Secuencia (6 Dígitos)
NINTER	N de Intervención
FECHAGEST	Fecha de la Gestión
MOTIVO	Motivo de la Gestión
IDGESTION	Identificador de la gestión
FECALTA	Fecha de alta de la gestión
CPROVI	Código de provincia
CCENTR	Código de distrito
CUNITS	Código de UTS
CODTIPGEST	Código del Tipo de la Gestión Realizada
CODRESPONSEXP	Código de responsable de expediente

TABLE XXII
TABLE TEQUEX

Name	Description
CEXCOM	Número de Expediente Completo: N Provincia(2) + N CSS (3) + N UTS(2) + N Secuencia (6 Dígitos)
CEQUIP	Código de Equipamiento
FECALTA	Fecha de alta

APPENDIX B
REMOVED COLUMNS

TABLE XXIII
REMOVED COLUMN FOR TABLE TEXPFA

Name	Type
CPROVI	Primary or foreign key / Code
CODCOA	Primary or foreign key / Code
DOMICI	Anonymized
DPOBLA	Anonymized
NTELEF	Anonymized
CEXCOM	Primary or foreign key / Code
PTEBORR	Primary or foreign key / Code
NTELEF2	Anonymized
NUMDOM	Anonymized
ENCRIP	Anonymized
CODPOS	Primary or foreign key / Code
INFORMEMOTIVO	Descriptive text
INFORMEDERIVADO	Descriptive text
INFORMECODPROFAL	Descriptive text
CODRESPONSEXP	Primary or foreign key / Code
DIAGNSOCIAL	Descriptive text
PRONOSTICO	Descriptive text
PROPS	Descriptive text
NUMEXPMUN	Primary or foreign key / Code
OBSERV	Descriptive text
SITECON	Descriptive text
INFORMESOLICITADO	Descriptive text
HISTFAM	Descriptive text
RELPADRESHIJOS	Descriptive text
VALORAPOYOOTRAS	Descriptive text
EQUIPA	Descriptive text
RELMUF	Descriptive text
ROLES	Descriptive text
ENFERMEDADES	Descriptive text
PERSCUIDMEN	Descriptive text
ESTABEMOC	Descriptive text
RELPAREJA	Descriptive text
LIMITACIONES	Descriptive text
NORMAS	Descriptive text
RELHERMANOS	Descriptive text
ABUSOS	Descriptive text
RELFAMESC	Descriptive text
RELNINOS	Descriptive text
ATNNECMED	Descriptive text
GRADOATN	Descriptive text
PLANFAM	Descriptive text

TABLE XXIV
REMOVED COLUMN FOR TABLE TMIEMB

Name	Type
CEXCOM	Primary or foreign key / Code
CINDFA	Primary or foreign key / Code
CTRMUN	Primary or foreign key / Code
DAPEL1	Anonymized
DAPEL2	Anonymized
DNOMBR	Anonymized
NUMDNI	Anonymized
NUMTARJSANIT	Anonymized
TITINFORME	Descriptive text
PASAPORTE	Anonymized
MEDICO	Descriptive text
CTRSAL	Descriptive text
HORCONS	Descriptive text
DIAGNDISCAP	Descriptive text
CURCENTRO	Descriptive text
SITHISTLAB	Descriptive text
DIRCTRSAL	Descriptive text
OBSERV	Descriptive text
NTELEF	Anonymized
TELCTRSAL	Anonymized
NTELEF2	Descriptive text

TABLE XXV
REMOVED COLUMN FOR TABLE TINTER

Name	Type
CEXCOM	Primary or foreign key / Code
NINTER	Primary or foreign key / Code
CODRESPONSINT	Primary or foreign key / Code
OBSER	Descriptive text

TABLE XXVI
REMOVED COLUMN FOR TABLE TUSUIN

Name	Type
CEXCOM	Primary or foreign key / Code
NINTER	Primary or foreign key / Code
CINDFA	Primary or foreign key / Code

TABLE XXVII
REMOVED COLUMN FOR TABLE TVALIN

Name	Type
CEXCOM	Primary or foreign key / Code
NINTER	Primary or foreign key / Code

TABLE XXVIII
REMOVED COLUMN FOR TABLE TDEMIN

Name	Type
CEXCOM	Primary or foreign key / Code
NINTER	Primary or foreign key / Code

TABLE XXIX
REMOVED COLUMN FOR TABLE TRECID

Name	Type
CEXCOM	Primary or foreign key / Code
NINTER	Primary or foreign key / Code

TABLE XXX
REMOVED COLUMN FOR TABLE TRECAP

Name	Type
CEXCOM	Primary or foreign key / Code
NINTER	Primary or foreign key / Code
CODCOM	Primary or foreign key / Code
DESTINO	Anonymized
NOMAPESPERCEP	Anonymized
DNIPERCEP	Anonymized
DOMPERCEP	Anonymized
LOCPERCEP	Anonymized
TELPERCEP	Anonymized
CODPROVPERCEP	Anonymized
RELPERCEPUSUS	Descriptive text
CINDFA	Primary or foreign key / Code
ENTIDAD	Descriptive text
OBSERV	Descriptive text
NUMCTAPERCEP	Anonymized
OBSERVALOJ	Descriptive text

TABLE XXXI
REMOVED COLUMN FOR TABLE TGESTINT

Name	Type
CEXCOM	Primary or foreign key / Code
NINTER	Primary or foreign key / Code
MOTIVO	Descriptive text
IDGESTION	Primary or foreign key / Code
CPROVI	Primary or foreign key / Code
CCENTR	Primary or foreign key / Code
CUNITS	Primary or foreign key / Code
CODRESPONSEXP	Primary or foreign key / Code

TABLE XXXII
REMOVED COLUMN FOR TABLE TEQUEX

Name	Type
CEXCOM	Primary or foreign key / Code

APPENDIX C
VALUES FOR TABLE TEQUEX

TABLE XXXIII
VALUES FOR TABLE TEQUEX

Code	Description	Column
1	AGUA CORRIENTE	TEQUEX_CEQUIP1
2	WC	TEQUEX_CEQUIP2
3	DUCHA	TEQUEX_CEQUIP3
4	ELECTRICIDAD	TEQUEX_CEQUIP4
5	GAS	TEQUEX_CEQUIP5
6	AGUA CALIENTE	TEQUEX_CEQUIP6
7	TELEFONO	TEQUEX_CEQUIP7
8	FRIGORIFICO	TEQUEX_CEQUIP8
9	CALEFACCION CASA ENTERA	TEQUEX_CEQUIP9
10	LAVADORA AUTOMATICA	TEQUEX_CEQUIP10
11	BARRERAS ARQUITECTONICAS EN EL ACCESO	TEQUEX_CEQUIP11
12	BARRERAS ARQUITECTONICAS EN LA VIVIENDA	TEQUEX_CEQUIP12
13	FALTA DE ILUMINACION NATURAL	TEQUEX_CEQUIP13
14	FALTA DE VENTILACION	TEQUEX_CEQUIP14
15	ESTADO DETERIORADO. GOTERAS/HUMEDAD	TEQUEX_CEQUIP15
16	AMENAZA DE RUINA	TEQUEX_CEQUIP16
17	ACEPTABLE	TEQUEX_CEQUIP17

APPENDIX D
H_2O.AI MODEL OUTPUT

```
ModelMetricsBinomial: drf
** Reported on test data. **

MSE: 0.0075820170936302805
RMSE: 0.08707477874580148
LogLoss: 0.037828635925550594
Mean Per-Class Error: 0.01245953520376153
AUC: 0.996795594245171
Gini: 0.9935911884903419
Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.16794199148813882:
       0     1    Error    Rate
-----  -----  ---  -------  --------------
0      20220  6    0.0003   (6.0/20226.0)
1      30     961  0.0303   (30.0/991.0)
Total  20250  967  0.0017   (36.0/21217.0)
Maximum Metrics: Maximum metrics at their respective thresholds

metric                          threshold    value     idx
------------------------------  -----------  --------  -----
max f1                          0.167942     0.981614  227
max f2                          0.147379     0.97593   231
max f0point5                    0.227752     0.990576  217
max accuracy                    0.167942     0.998303  227
max precision                   0.997054     1         0
max recall                      0.000248553  1         398
max specificity                 0.997054     1         0
max absolute_mcc                0.167942     0.980806  227
max min_per_class_accuracy      0.0759326    0.984864  263
max mean_per_class_accuracy     0.108639     0.98754   243
Gains/Lift Table: Avg response rate:  4.67 %
```

	group	cumulative_data_fraction	lower_threshold	lift	cumulative_lift	response_rate	cumulative_response_rate	capture_rate	cumulative_capture_rate	gain	cumulative_gain
	1	0.0101334	0.874518	21.4097	21.4097	1	1	0.216953	0.216953	2040.97	2040.97
	2	0.0200311	0.763422	21.4097	21.4097	1	1	0.211907	0.42886	2040.97	2040.97
	3	0.0300231	0.630521	21.4097	21.4097	1	1	0.213925	0.642785	2040.97	2040.97
	4	0.0400151	0.436528	21.4097	21.4097	1	1	0.213925	0.85671	2040.97	2040.97
	5	0.0500071	0.100217	12.2197	19.5734	0.570755	0.914232	0.122099	0.978809	1121.97	1857.34
	6	0.100014	0.0498971	0.242145	9.90778	0.0113101	0.462771	0.012109	0.990918	-75.7855	890.778
	7	0.150021	0.0375886	0.0403576	6.61864	0.00188501	0.309142	0.00201816	0.992936	-95.9642	561.864
	8	0.200028	0.0305685	0.0605363	4.97911	0.00282752	0.232564	0.00302725	0.995964	-93.9464	397.911
	9	0.299995	0.0210738	0	3.31993	0	0.155067	0	0.995964	-100	231.993
	10	0.400009	0.0147458	0.0100894	2.49237	0.000471254	0.116413	0.00100908	0.996973	-98.9911	149.237
	11	0.500024	0.0107788	0	1.99385	0	0.0931285	0	0.996973	-100	99.3852
	12	0.599991	0.00668682	0.0201883	1.66501	0.000942951	0.077769	0.00201816	0.998991	-97.9812	66.5011
	13	0.700005	0.00348336	0	1.42712	0	0.0666577	0	0.998991	-100	42.712
	14	0.799972	0.00172864	0	1.24878	0	0.0583279	0	0.998991	-100	24.8783
	15	0.899986	0.000711206	0	1.11001	0	0.051846	0	0.998991	-100	11.0007
	16	1	0	0.0100894	1	0.000471254	0.0467078	0.00100908	1	-98.9911	0

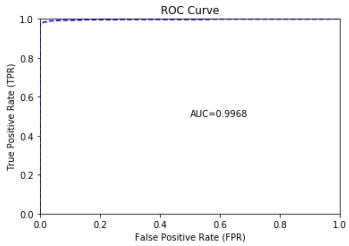


Fig. 37. H_2O.ai model output

APPENDIX E
PARAMETERS USED IN THE MACHINE LEARNING MODELS

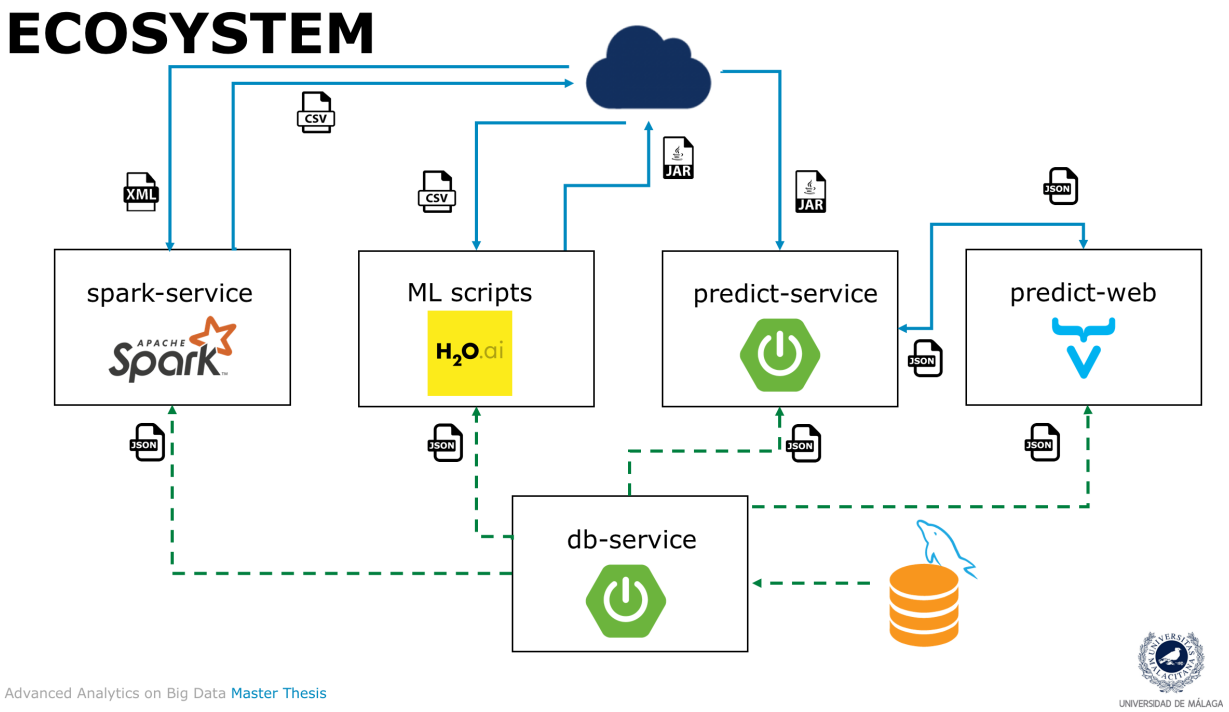


Fig. 38. Ecosystem diagram

APPENDIX F
PARAMETERS USED IN THE MACHINE LEARNING MODELS

TABLE XXXIV
PARAMETER USED

Parameter	Description	Models
nfolds	This option specifies the number of folds to use for cross-validation	ALL
ignore_const_cols	Specify whether to ignore constant training columns, since no information can be gained from them	ALL
stopping_metric	Specify the metric to use for early stopping. - logloss: The confidence assigned to the correct category is used to calculate logloss. Logloss disproportionately punished low numbers, which is another way of saying having high confidence in the wrong answer is a bad thing - misclassification: This is overall error	RF/DL
stopping_rounds	Stops training when the option selected for stopping_metric doesnt improve for the specified number of training rounds, based on a simple moving average	RF/DL
stopping_tolerance	Specify the relative tolerance for the metric-based stopping to stop training if the improvement is less than this value	RF/DL
fold_assignment	(Applicable only if a value for nfolds is specified and fold_column is not specified) Specify the cross-validation fold assignment scheme. The available options are AUTO (which is Random), Random,Modulo, or Stratified (which will stratify the folds based on the response variable for classification problems)	RF
balance_classes	(Applicable for classification only) Specify whether to oversample the minority classes to balance the class distribution	RF
ntrees	Specify the number of trees	RF
max_depth	Specify the maximum tree depth	RF
min_rows	Specify the minimum number of observations for a leaf	RF
epochs	Specify the number of times to iterate (stream) the dataset. The value can be a fraction.	DL
train_samples_per_iteration	Specify the number of global training samples per MapReduce iteration. To specify one epoch, enter 0. To specify all available data (e.g., replicated training data), enter -1. To use the automatic values, enter -2.	DL
family	Specify the model type. If the family is multinomial, the data can be categorical with more than two levels/classes (Enum)	GLM

ACKNOWLEDGMENT

First of all, I would like to thank to the workers of Observatorio Municipal para la Iclusión Social of Málaga City Council, Lola, Paco and Juan A., for their kindness with me, their time and their readiness for giving me the data with whom this Project has been developed and all information that I needed to understand their day to day work. From my meetings with them has resulted the idea of this Project. I would also like to thank to Luis Gómez Jacinto, PhD in Psychology of University of Málaga for his idea of introduce the Big Data and the Data Analytics in the social science scope, in the social work, particularly, where, honestly, I think that there is a growth line very interesting.

Of course, I thank to my family and Teresa, for their supported and their patience for the last one year and a half where they have had suffer my brainstorms, becoming in the best rubber ducks and to my workers mate for their tips. Also, to mi dog Sansa and my cat Robin for their company.

Finally, I would like to give a very special thanks to all social workers who devoting their day to day to helping people, trying to build a more just and egalitarian society. I hope that I was able to contribute my bit.

REFERENCES

[1] Observatorio Municipal para la Inclusión Social. Ayuntamiento de Málaga, http://observatoriosocial.malaga.eu/

[2] Ministerio de Sanidad, Servicios Sociales e Igualdad. Gobierno de España, http://www.msssi.gob.es/ssi/portada/home.htm

[3] H_2O.ai, https://www.h2o.ai/

[4] Darren Cook, *Practical Machine Learning with H2O*, 1st ed. O'Really, 2017.

[5] Observatorio Municipal para la Inclusión Social, *Perfil de las personas usuarias de los SSAP de Málaga, 2013*, 2014.

[6] Apache Spark, https://spark.apache.org/docs/latest/