

Credit Card (Binary Classification)

The goal of this study is to train a model in order to predict whether the application is Approved (0) or Rejected (1). The dataset used in this case study is found in <https://www.kaggle.com/datasets/rohitudageri/credit-card-details?select=Credit_card_label.csv> and has 20 features and 1458 labelled samples.

# Step 1: Import Data from the file

Right click on the input spreadsheet and choose the option “Import from file”. Then navigate through your files to find the one with the credit card data.

A screenshot of a spreadsheet

Description automatically generated

A screenshot of a computer

Description automatically generated

# Step 2: Manipulate Data

In order to use the data for training we have to exclude any columns that do not represent factor, like Ind\_ID. We follow these steps to execute this:

* Browse: “Data Transformation” → “Data Manipulation” → “Select Column(s)”.
* Select all columns except the one that corresponds to the Ind\_ID.

|  |  |
| --- | --- |
| A screenshot of a computer  Description automatically generated | A screenshot of a computer  Description automatically generated |

The data without the Ind\_ID column will appear in the output spreadsheet.

# Step 3: Fill missing values

There are empty values in the Dataset. Specifically, we show below how many missing values there are for each feature:

A screenshot of a computer screen

Description automatically generated

Create a new tab by pressing the “+” button on the bottom of the page with the name FILL\_MISSING VALUES which will be used to fill the missing values.

Import Data into the input spreadsheet of the FILL\_MISSING\_VALUES tab from the output of the IMPORT tab by right-clicking on the input spreadsheet and then choosing Import from SpreadSheet.

A screenshot of a computer

Description automatically generated

Handle missing columns values by browsing: “Data Transformation” → “Data Manipulation” → “Fill missing column(s) Values”. Then choose the Mean as the Numerical Method.

|  |  |
| --- | --- |
|  | A screenshot of a computer  Description automatically generated |

The results will appear on the output spreadsheet.

A screenshot of a computer

Description automatically generated

# Step 4: Split Data

Create a new tab by pressing the + button on the bottom of the page with the name TRAIN\_TEST\_SPLIT which we will use for splitting to create the train and test set.

Import Data into the input spreadsheet of the TRAIN\_TEST\_SPLIT tab from the output of the FILL\_MISSING\_VALUES tab by right-clicking on the input spreadsheet and then choosing Import from SpreadSheet.

A screenshot of a computer

Description automatically generated

Split the dataset by browsing “Data Transformation” → “Split” → “Random Partitioning”. Then choose the training set percentage and the column for the sampling as shown below.

|  |  |
| --- | --- |
| A screenshot of a computer  Description automatically generated | A screenshot of a computer  Description automatically generated |

The results will appear on the output spreadsheet.

A screenshot of a computer

Description automatically generated

# Step 5: Normalize the Training Set

Create a new tab by pressing the + button on the bottom of the page with the name NORMALISE\_TRAIN\_SET.

Import Data into the input spreadsheet of the NORMALISE\_TRAIN\_SET tab the train set from the output of the TRAIN\_TEST\_SPLIT tab by right-clicking on the input spreadsheet and then choosing Import from SpreadSheet. From the available Select input tab options choose TRAIN\_TEST\_SPLIT: Training Set

A screenshot of a computer

Description automatically generated

Normalize the Data using Z-score by browsing: “Data Transformation” → “Normalize” → “Z-Score”. Then select all columns excluding Mobile\_phone and Label and click Execute.

|  |  |
| --- | --- |
| A screenshot of a computer  Description automatically generated | A screenshot of a computer  Description automatically generated |

Τhe results will appear on the output spreadsheet.

A screenshot of a computer

Description automatically generated

# Step 6: Normalize the Test Set

Create a new tab by pressing the + button on the bottom of the page with the name NORMALISE\_TEST\_SET.

Import Data into the input spreadsheet of the NORMALISE\_TEST\_SET tab the test set from the output of the TRAIN\_TEST\_SPLIT tab by right-clicking on the input spreadsheet and then choosing Import from SpreadSheet. From the available Select input tab options choose TRAIN\_TEST\_SPLIT: Test Set.

A screenshot of a computer

Description automatically generated

Normalize the test set using the existing normalizer of the training set by browsing: “Analytics” → “Existing Model Utilization” → “Model: NORMALIZE\_TRAIN\_SET”.

|  |  |
| --- | --- |
|  | A screenshot of a computer  Description automatically generated |

The results will appear on the output spreadsheet.

A screenshot of a computer

Description automatically generated

# Step 7: Train the model

Create a new tab by pressing the “+” button on the bottom of the page with the name “TRAIN\_MODEL(.fit)”.

Import data into the input spreadsheet of the “TRAIN\_MODEL(.fit)” tab from the output of the “NORMALISE\_TRAIN\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from SpreadSheet”.

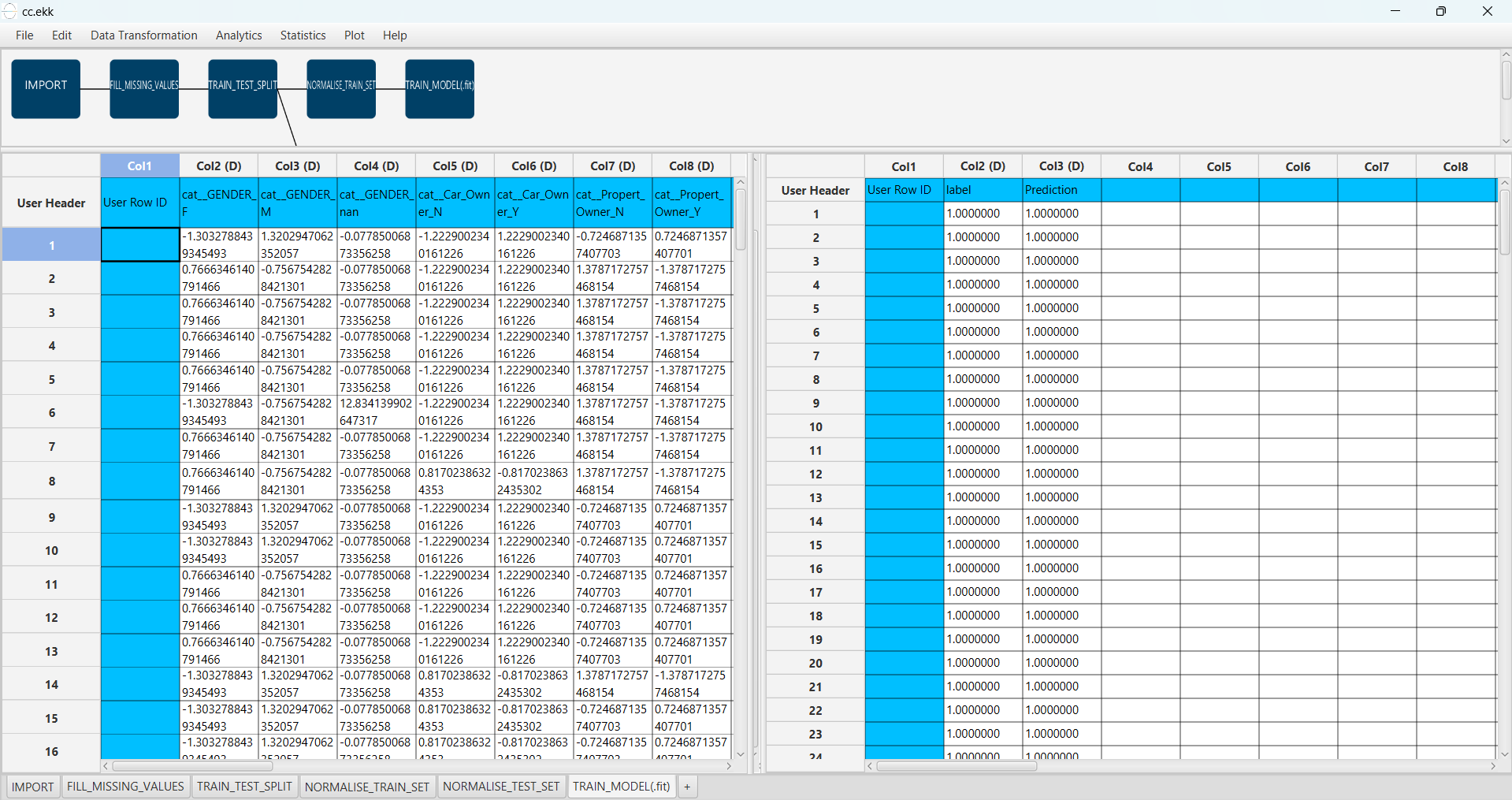
A screenshot of a computer

Description automatically generated

Use the XGBoost Method to train and fit the model by browsing: “Analytics” → “Classification” → “XGBoost” and set the “number of estimators” as 210, the “column sample by tree” as 10, the “Target Column” as the column corresponding to “Label” and use the following “RNG Seed”: 1732285527644.

|  |  |
| --- | --- |
| A screenshot of a computer  Description automatically generated | A screenshot of a computer  Description automatically generated |

The predictions will appear on the output spreadsheet.



# Step 8: Validate the model

Create a new tab by pressing the “+” button on the bottom of the page with the name “VALIDATE\_MODEL(.predict)”.

Import data into the input spreadsheet of the “VALIDATE\_MODEL(.predict)” tab from the output of the “NORMALISE \_TEST\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from SpreadSheet”.

To validate the model: “Analytics” → “Existing Model Utilization”. Then choose Model “(from Tab:) TRAIN\_MODEL (.fit)”. and transfer the “label” column to the output.

|  |  |
| --- | --- |
| A screenshot of a computer  Description automatically generated | A screenshot of a computer  Description automatically generated |

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedThe predictions will appear on the output spreadsheet.

# Step 9: Statistics calculation

Create a new tab by pressing the “+” button on the bottom of the page with the name “STATISTICS\_ACCURACIES”.

Import data into the input spreadsheet of the “STATISTICS\_ACCURACIES” tab from the output of the “VALIDATE\_MODEL(.predict)” tab by right-clicking on the input spreadsheet and then choosing “Import from SpreadSheet”.

Calculate the statistical metrics for the classification by browsing: “Statistics” → “Model Metrics” → “Classification Metrics”.

|  |  |
| --- | --- |
| A screenshot of a computer  Description automatically generated |  |

The results will appear on the output spreadsheet.

Accuracy: 0.902

F1-Score = 0.694

A screenshot of a computer

Description automatically generated

# Step 10: Reliability check of each record of the test set

## Step 10.a: Create the domain

Create a new tab by pressing the “+” button on the bottom of the page with the name “EXCLUDE\_LABEL”.

A screenshot of a computer

Description automatically generatedImport data into the input spreadsheet of the “EXCLUDE\_LABEL” tab from the output of the “NORMALISE\_TRAIN\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from SpreadSheet”.

Manipulate the data to exclude the column that corresponds to the “label” by browsing: “Data Transformation” → “Data Manipulation” → “Select Columns”. Then select all the columns except the “label”.

|  |  |
| --- | --- |
| A screenshot of a computer  Description automatically generated | A screenshot of a computer  Description automatically generated |

The results will appear on the output spreadsheet.

Create a new tab by pressing the “+” button on the bottom of the page with the name “DOMAIN”.

Import data into the input spreadsheet of the “DOMAIN” tab from the output of the “EXCLUDE\_LABEL” tab by right-clicking on the input spreadsheet and then choosing “Import from SpreadSheet”.

A screenshot of a computer

Description automatically generated

Create the domain by browsing: “Statistics” → “Domain APD”.

|  |  |
| --- | --- |
| A screenshot of a computer  Description automatically generated | A screenshot of a computer  Description automatically generated |

The results will appear on the output spreadsheet.

A screenshot of a computer

Description automatically generated

## Step 10.b: Check the test set reliability

Create a new tab by pressing the “+” button on the bottom of the page with the name “EXCLUDE\_LABEL\_TEST\_SET”.

Import data into the input spreadsheet of the “EXCLUDE\_LABEL\_TEST\_SET” tab from the output of the “NORMALISE \_TEST\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from SpreadSheet”.

A screenshot of a computer

Description automatically generated

Filter the data to exclude the column that corresponds to the “label” by browsing: “Data Transformation” → “Data Manipulation” → “Select Columns”. Then select all the columns except “label”.

|  |  |
| --- | --- |
| A screenshot of a computer  Description automatically generated | A screenshot of a computer  Description automatically generated |

The results will appear on the output spreadsheet.

Create a new tab by pressing the “+” button on the bottom of the page with the name “RELIABILITY”.

Import data into the input spreadsheet of the “RELIABILITY” tab from the output of the “EXCLUDE\_LABEL\_TEST\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from SpreadSheet”.

A screenshot of a computer

Description automatically generated

Check the Reliability by browsing: “Analytics” → “Existing Model Utilization”. Then select as Model “(from Tab:) DOMAIN”.

|  |  |
| --- | --- |
| A screenshot of a computer  Description automatically generated | A screenshot of a computer  Description automatically generated |

The results will appear on the output spreadsheet.

A screenshot of a computer

Description automatically generated

There are four unreliable samples in the test set.

# Final Isalos Workflow

Following the above-described steps, the final workflow on Isalos will look like this:

A diagram of a model

Description automatically generated with medium confidence