**ENHANCING AUDITOR DECISION-MAKING IN INDIAN FIRMS WITH A PREDICTION CLASSIFICATION MODEL**

**Team:**

1. Srikanth Ranganathan (Student Number: 1009747794)

Contact: [srikanth.ranganathan@mail.utoronto.ca](mailto:srikanth.ranganathan@mail.utoronto.ca)

1. Joshua Samadi (Student Number: 1005627403)

Contact: [josh.samadi@mail.utoronto.ca](mailto:josh.samadi@mail.utoronto.ca)

1. **Explain what problem you are going to solve using this dataset. Provide a brief overview of your problem statement.**

We aim to address the challenge of identifying potential instances of fraudulent financial activities among firms audited by the Auditor Office in India during the years 2015 and 2016. Leveraging a dataset comprising 776 data points, we intend to construct a predictive classification model. This model will utilize 26 input variables, including Sector, Sector\_Score, Location, money\_value, district loss, and various Risk types, among others. By deploying this predictive model, auditors can make informed decisions during the auditing process, gaining valuable insights into whether a particular firm may exhibit signs of suspicious financial behavior. The classification model will serve as a practical tool for auditors, offering a reference point to assess and identify potentially fraudulent activities within audited firms based on the provided input variables.

1. **Explain your dataset. Explore your dataset and provide at least 5 meaningful charts/graphs with an explanation.**
2. **Do data cleaning/pre-processing as required and explain what you have done for your dataset and why.**

The audit\_risk dataset has 776 rows and 27 columns. There are 26 feature columns and 1 target column (Risk)

A group of people on a white background

Description automatically generated

Below table provides statistical summary for all the features and target variables.

A screenshot of a computer screen

Description automatically generated

Below table provides features type for each column. The dataset has 23 column of double-precision Float64 real number, 3 columns of 64-bit integer data and 1 object column.

A screenshot of a computer program

Description automatically generated

As per following table there is only one null value in Money\_Value column

A screenshot of a computer

Description automatically generated

Below Histogram charts demonstrate distribution type across different features. it is observed that most of the features has non-normal distribution with some features distribution scattered across the range.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

In order to get more insight from some of the features, we plotted cross table between some features and target column (“Risk”). The first feature is Sector\_Score which sector score 2.37, 2.72, 3.41 and 3.89 sectors has demonstrated the highest amount of being fraudulent.

A screenshot of a computer screen

Description automatically generated A graph of different colored bars

Description automatically generated

For District\_Loss, District #2 has the highest amount of being fraudulent firm.

A screenshot of a graph

Description automatically generated A graph of a loss

Description automatically generated with medium confidence

For Location\_ID, Locations #8, #19 and #2 have this highest amount of fradulant firms.

A screenshot of a computer

Description automatically generated

Below boxplot demonstrate the spread and skewness of different features for dataset

A screen shot of a graph

Description automatically generated A diagram of a number of blue boxes

Description automatically generated

A graph with a graph and a graph with a graph and a graph with a graph and a graph with a graph and a graph with a graph and a graph with a graph and a graph with

Description automatically generated A graph of a graph with blue squares

Description automatically generated

A graph of a risk

Description automatically generated A graph with a bar graph and a graph with a bar graph

Description automatically generated with medium confidence

A blue square with white squares

Description automatically generated A graph with a blue bar and black lines

Description automatically generated with medium confidence

A graph with a bar and a number of points

Description automatically generated with medium confidence A graph with numbers and points

Description automatically generated

A graph with a blue line and a blue line

Description automatically generated A blue square with a black dot

Description automatically generated

A graph with a blue line and a blue line

Description automatically generated A graph of a loss by target variable

Description automatically generated

A graph with a blue square and a black line

Description automatically generated A graph of a graph with a bar chart and a graph with a bar chart

Description automatically generated with medium confidence

A graph with a blue bar and black dots

Description automatically generated with medium confidence A graph with a blue square and a blue square

Description automatically generated

Pairplot and correlation matrix below demonstrated strong correlation between following features

* Risk\_A & PARA\_A
* Risk\_B & TOTAL
* TOTAL & PARA\_B
* Score\_B.1 & Risk\_C
* Risk\_C & Number
* Risk\_B.1 & Number
* Risk\_D & Money Value
* Risk\_E & District\_Loss
* Prob & history
* History & Risk\_F
* Risk\_B and Score

A screenshot of a graph

Description automatically generated

A collage of a grid of squares

Description automatically generated

A group of squares with orange and blue dots

Description automatically generated

A screenshot of a computer

Description automatically generated

# In order to clean the dataset, rows that contain objects are imputed with number.

# correlated features are getting dropped from the dataset to avoid complexity of the algorithm and increasing the risk of errors. New dataset with 14 feature and 1 target columns will be used for machine learning purpose.

A screenshot of a computer

Description automatically generated

**Q5) USING AUTOMATED ML ON AZURE ML STDUIO ON THE DATASET**A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

It can be noted that the LightGBM (Light Gradient Boosting Machine) has been concluded as the best model.

A screenshot of a computer

Description automatically generated

It is very clear as to why this model has been concluded as the best model, it is because of the resulting metric scores such as accuracy, precision, recall, and f1 score, all of them being 1.

The FP and FN rates are also 0.

Here are some plots that further validate this:

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated