

# DATA WAREHOUSE AND DATA MINING



## BANKING DATA ANALYSIS [Detailed Analysis of a Credit Score Card Predictor] FINAL REPORT

**Submitted By:**

Anukriti (2K17/IT/027)

Anurag Mudgil (2K17/IT/029)

## LEHMAN BROTHERS BANKRUPTCY (2008)

On September 15, 2008, Lehman Brothers Holdings, Inc. initiated the largest bankruptcy proceeding in United States history. It declared \$639 billion in assets and \$613 billion in debts. At the time, Lehman was the fourth-largest U.S. investment bank, with 25,000 employees worldwide, a far cry from its humble beginnings in 1844 in Montgomery, Alabama, as a drygoods store.



Despite being thought ‘too big to fail’, the federal government did not employ extraordinary measures to save Lehman. Lehman's demise was a seminal event in the financial crisis that began in the U.S. subprime mortgage industry in 2007, spread to the credit markets, and then burned through the world's financial markets. The crisis resulted in significant and wide losses to the economy. One cause of Lehman's demise was its significant exposure to the U.S. subprime mortgage and real estate markets.

Lehman Brothers experienced a credit crunch. With no loans being made and the world's largest financial institutions under significant threat of failure, the global financial system was under threat of collapse.

### BUSINESS PROBLEM:

Lehman, like most investment banks, relied on these short-term markets to raise billions of dollars each day. Ultimately, it was an inability to secure funding that was Lehman's undoing. Other factors contributing to Lehman's failure were a highly-leveraged, risk-taking business strategy supported by limited equity; a culture of excessive risk-taking. This indicates how important it is for the financial institutions to use a credit scorecard mechanism. Credit scoring is a dependable assessment of a person's credit worthiness since it is based on actual data. With the help of this, they will be able to make two significant decisions: first, whether to grant credit to a new applicant, and second, how to deal with existing applicants, including whether to increase their credit limits.

## STUDY OF VARIOUS CREDIT SCORING MODELS

A credit score reflects the likelihood that a consumer will repay his debts. With so many scoring methods used to determine your credit score, the variety of models means your score can vary several points depending on whose model is used and what type of business is asking for it.

- ❖ **FICO Score:** The FICO scoring model is considered the most reliable because it has the best track record. Regardless of which FICO model is used, there are five factors that mostly influence a classic FICO score and help to define your credit score: Payment history, Credit utilization, Credit history, Types of credit and New credit. There are many sub-categories calculated within each area before arriving at a final score.
- ❖ **Vantage Score:** The VantageScore model looks at familiar data, things like paying on time, keeping credit card balances low, avoiding new credit obligations, bank accounts and other assets to calculate its score. A person who is paying down debt is now likely to be scored better than a person who is making minimum payments and slowly accumulating credit card debt. The VantageScore uses information from all three credit reporting bureaus, but weighs certain factors more heavily or less heavily than the FICO algorithm. Thus, the scores should be similar, but rarely identical.
- ❖ **TransRisk:** It's based on data from TransUnion and determines an individual's risk on new accounts, instead of existing accounts. Because of that specialized nature, there's not much information available about the TransRisk score. Accordingly, it isn't utilized by many lenders. It has been reported that an individual's TransRisk score has generally been drastically lower than their FICO score.
- ❖ **CreditXpert Credit Score:** It was developed to help businesses approve new account candidates. It inspects credit reports for ways to raise its score quickly or detect false information. By improving those scores, that should lead to more loan approval for customers.

**Our Approach:** In order to determine the risk in lending, we decided to propose our own credit score card predictor. Through our project, we will be providing detailed analysis on how the actual scorecard predictor operates along with several decisions and implications while following the methodology. Our project has various application areas in the consumer market including credit cards, auto loans, home mortgages, home equity loans, mail catalog orders, and a wide variety of personal loan products.

## OUR METHODOLOGY



After going through various credit scorecard models we came up with our own scorecard model. We initially tried to figure out ways to find a suitable dataset for our problem but the data found required a lot of pre-processing. The dataset was collected from a bank and for each such record, the customer was given a status - good or bad.

We used the following features: age, gender, seniority, marital status, job, incomes/salary, housing (rent, own, for free), geographical (urban/rural), residential status, existing client (Y/N), number of years as client, total debt, account balance etc. These were around 13 features in total, out of which we

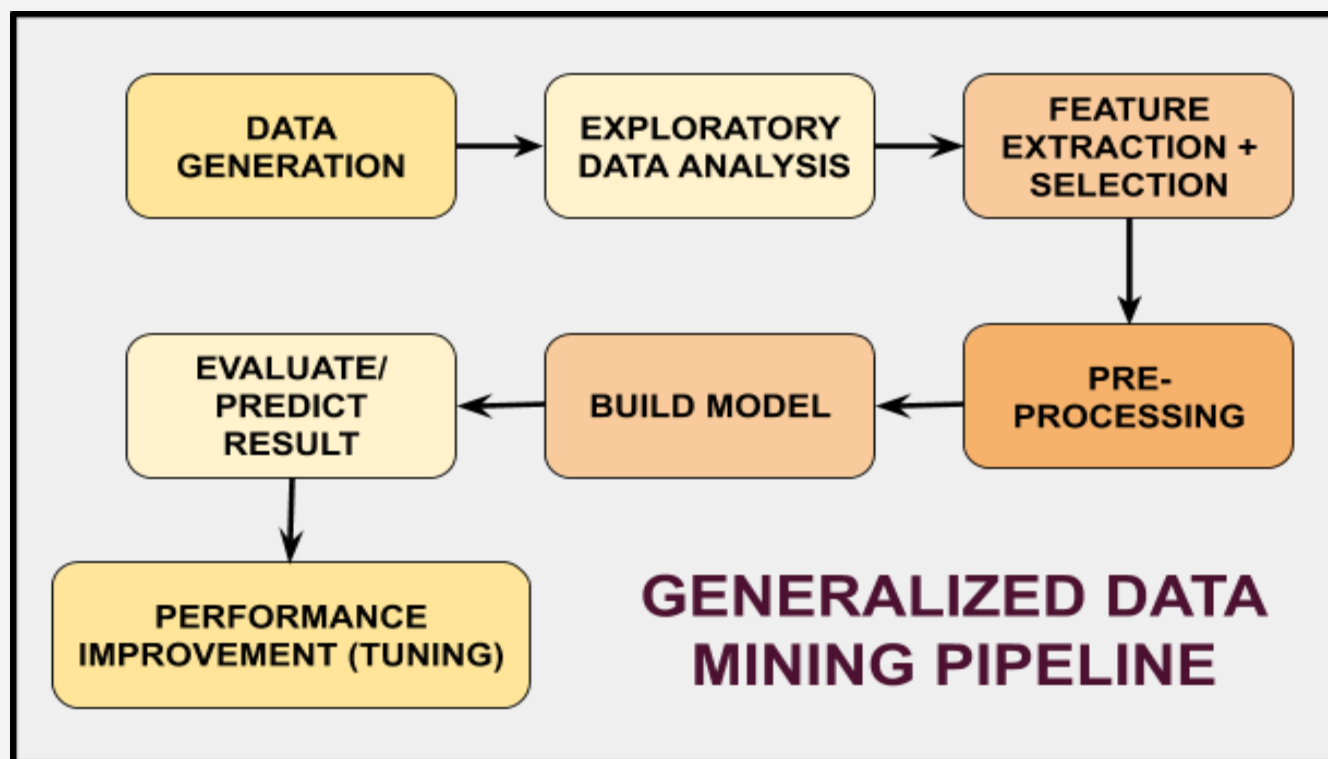
chose 8 most important and reliable features and trained our model on them. Two derived attributes were added later to increase the accuracy of the model.

Different types of variables used:

- **Categorical features:** These features have categories (Age of the customer, Status of the customer, Type of job, Records, Marital Status, Type of home ownership)
- **Numerical features:** These features have numerical values (Expenses, Income, Assets, Debt, Amount, Price)

We then performed exploratory data analysis and checked for missing values, nature of variables, interactions between useful variables, impact of variables on the target variables, obtained correlations, outlier analysis, feature importances etc. On the basis of this analysis, we performed the required preprocessing. Then for the model, we selected the Random Forest Model as our baseline model and tried to enhance their performance using hyper-parameter tuning employing Random forest CV. We could successfully reach the state-of-the-art accuracy of around 98.78%. Later, for convenience, we deployed our model using flask and built a GUI for demo purposes that can be easily used by any financial institution to calculate the credibility of their customers and hence, save them from any financial debt situations. Now, we will be explaining this entire methodology step-by-step indicating all our decisions and choices made during this process.

This project follows a general pipeline as shown below:



## STEP 1 (Defining Problem Statement):

We can consider this problem of detecting good/bad customers depending upon their credit score or risk associated as a binary classification task and the percentage associated with the label will determine the credit worthiness. We can therefore set a threshold and if a customer has worthiness less than the threshold, then he/she should not be provided with any loan as they are associated with high risk. This is a typical supervised learning task to predict into either of the two labels given the features.

## STEP 2 (Relevant Data Collection):



It was important to find a relevant training dataset and luckily, we found this dataset- [Mineria de Dades Credit](#). This dataset was well researched and benchmarked by the relevant community of researchers. But this data required a lot of pre-processing and analysis after which we created our customized dataset.

## Description of the dataset:

The dataset was collected from a bank and it contains 4446 customer records and for each such record, the customer was given a Status - good or bad depending upon 13 other customer related attributes which are given as follows: Seniority: job seniority in years, Home: type of home owned, Time: time of requested loan, Age: client's age, Marital: marital status, Records: existence of records, Job: type of job, Expenses: amount of expenses, Income: amount of income, Assets: amount of assets, Debt: amount of debt, Amount: amount requested for loan, Price: price of the good.

**Status {good, bad}:** Each customer in the dataset is labeled by good if he/she is credit worthy else bad.

## Addition of derived attributes:

- **Financial Ratio:** We will be calculating debt ratio which is used to measure the firm's ability to repay long-term debt. Ratios can be expressed as a decimal or an equivalent percent value. It is given by  $\text{Amount/Price}$ .
- **Savings:** In order to be debt-free, savings are a must and hence, becomes an important consideration for calculating credit worthiness of a customer. It is given by  $(\text{Income} - (\text{Expenses} + \text{Debt})) / (\text{Amount/Time})$ .

## STEP 3 (Exploratory Data Analysis) and

## STEP 4 (Pre-processing and Feature Selection):

### General observations and checking missing values:

- Income has 31 missing values.
- Assets have 41 missing values.
- Debt has 12 missing values.

### Required preprocessing:

These missing values are replaced by the mean values of their corresponding column.


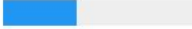
Status	True
Home	True
Marital	True
Records	True
Job	True

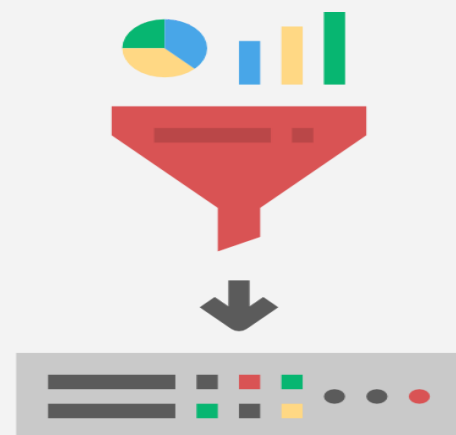
**Checking nature of variables:** There are 5 categorical variables as per the analysis as shown:



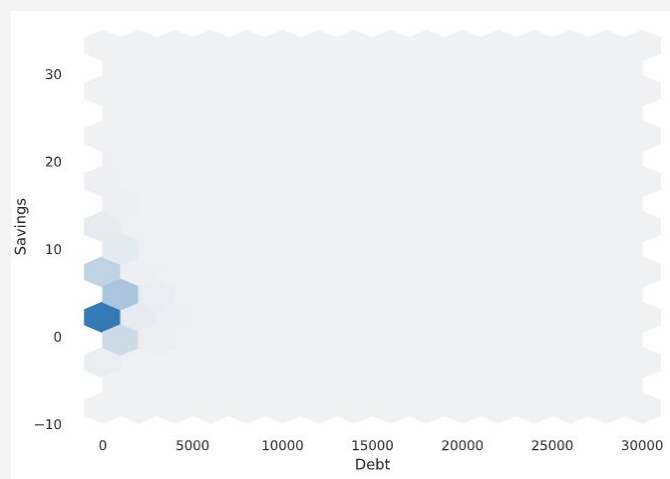
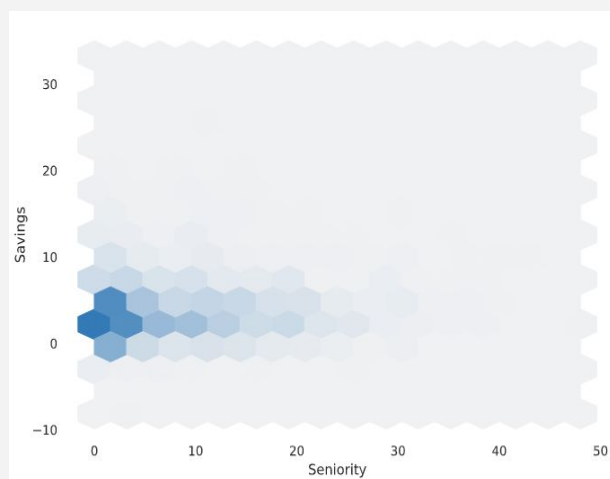
**Required preprocessing:** LabelEncoder was used to handle all the categorical variables and this converted the data into usable form.

**Distribution of the classes in the dataset:** It reveals if the dataset is an imbalanced dataset or not and if sophisticated techniques like SMOTE, ADASYN etc are required to handle data imbalance. The data found is imbalanced.

Distinct	2	1		3197
Distinct (%)	< 0.1%	0		1249
Missing	0			
Missing (%)	0.0%			



**Studying useful variable interactions:**

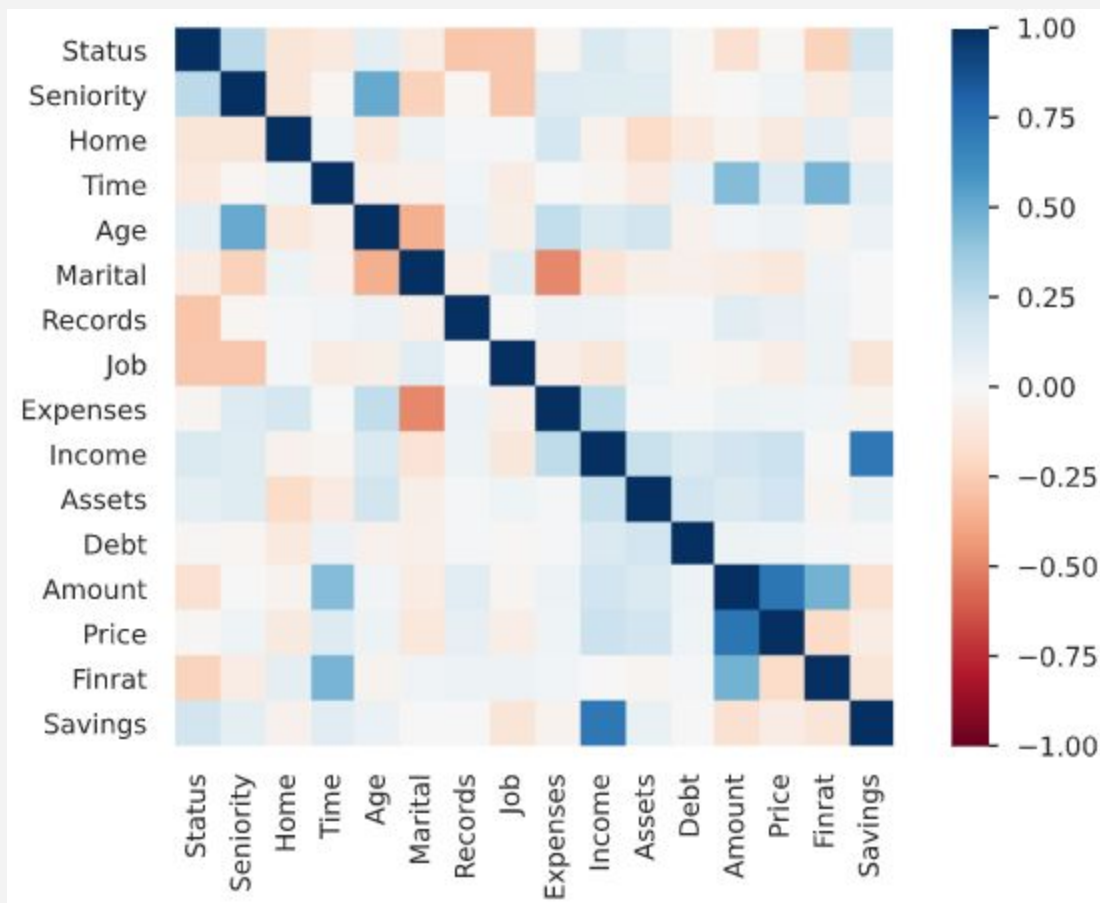


**Correlation Study:** We used Pearson's correlation coefficient(r) which is a measure of the strength of the association between the two variables. This analysis is important to figure out the most relevant features with respect to the target variable.

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y}) / (n - 1)}{\hat{\sigma}_X \hat{\sigma}_Y}.$$

We also obtained the top 10 and least 5 correlated variables which can be observed below:





```
Status      1.000000
Seniority    0.272161
Savings      0.210706
Assets       0.177243
Income       0.161955
Age          0.094434
Price        0.004481
Debt         0.004058
Expenses     -0.019253
Marital      -0.075793
Name: Status, dtype: float64

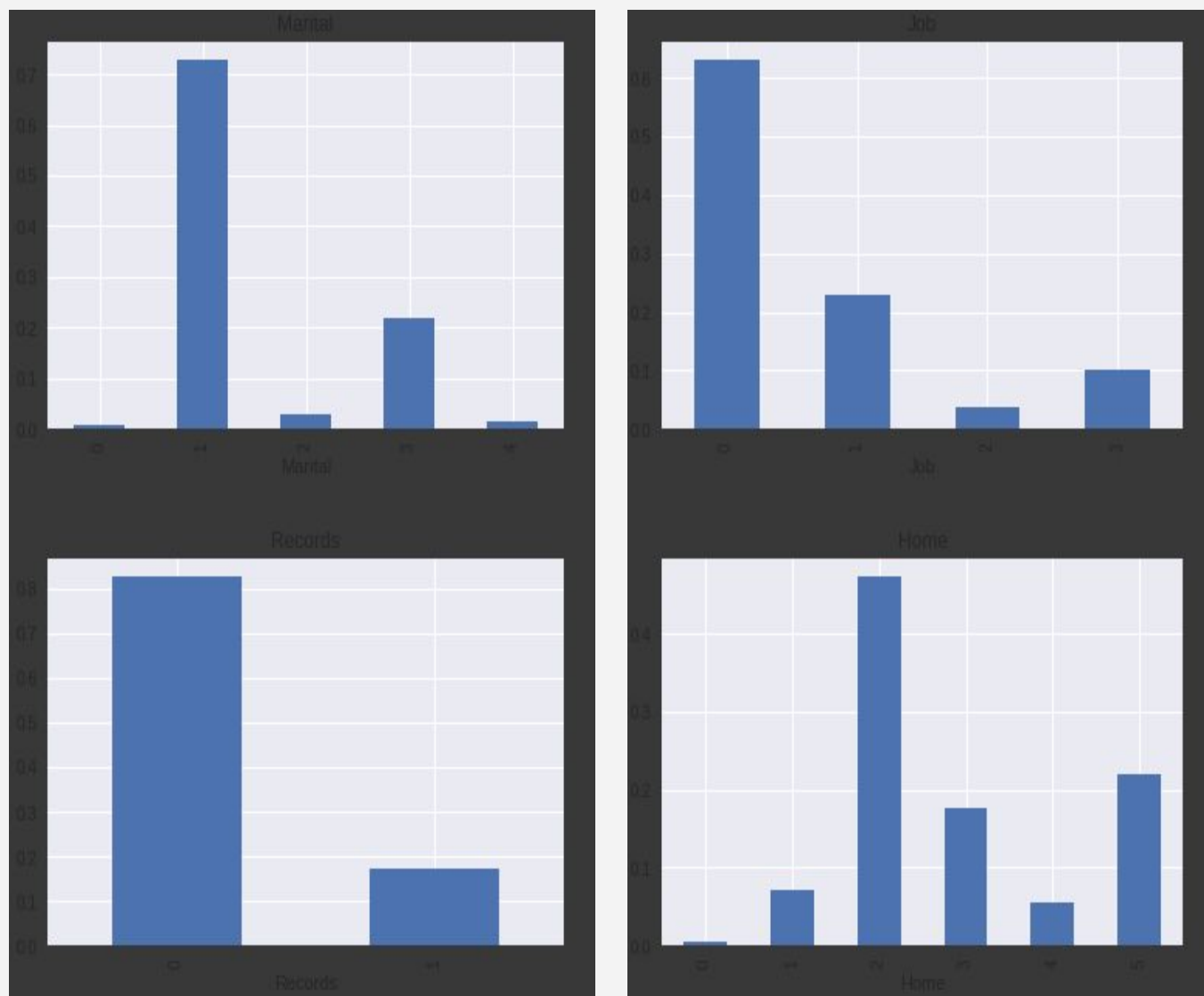
Home        -0.123475
Amount      -0.147990
Finrat      -0.221601
Job         -0.267701
Records     -0.277817
Name: Status, dtype: float64
```

### Bi-variate Analysis:

The t-test tells how significant the differences between groups are. Every t-value has a p-value to go with it. A p-value is the probability that the results from your sample data occurred by chance. They are usually written as a decimal. Low p-values are good as they indicate that data did not occur by chance. So, we calculated t-stats and p-value for each of the variables for better analysis.



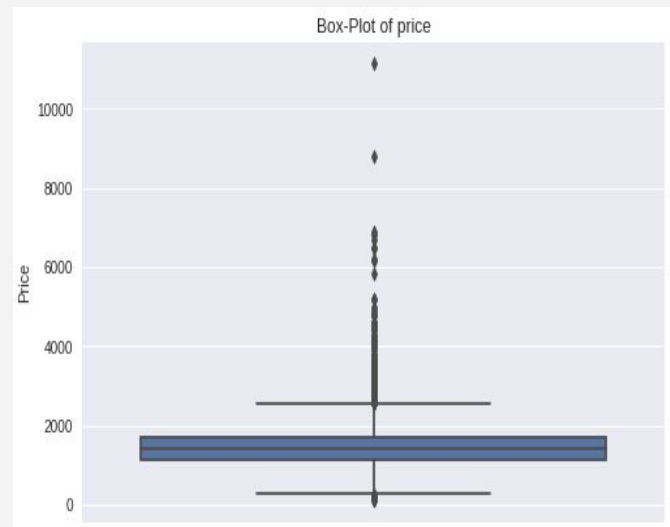
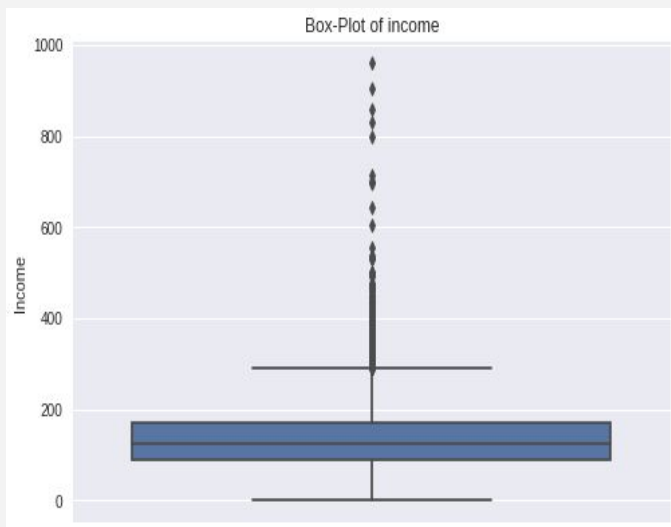
**Percentage of observations per label:** There are 5 categorical variables out of which 4 are features and the 5th one is the target variable. Histograms are obtained to analyze the count of each category corresponding to a particular label. This is shown below:



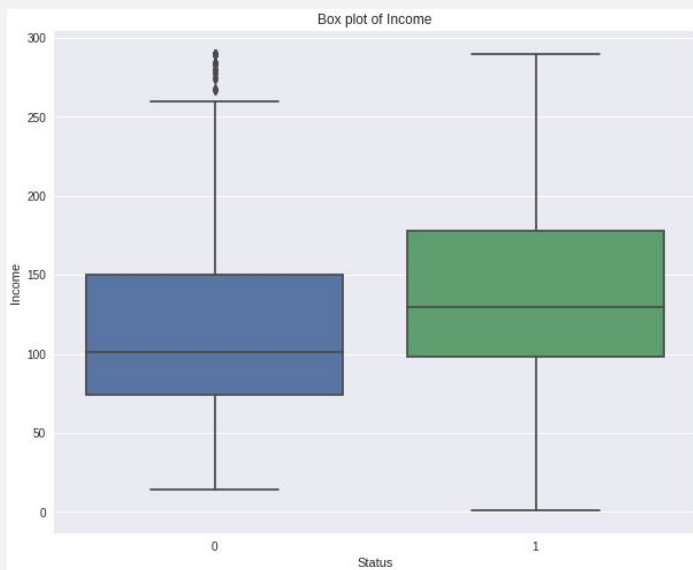
**Outlier Analysis:** Box plot study was done in order to observe the outliers present in each variable. Outliers were found mainly in 4 variables: 'Assets', 'Finrat', 'Price', 'Savings'.

**Required preprocessing:** Their skewness values were calculated based on which outlier capping methodology was adopted in which above 95th quantile, all the other values were clipped.

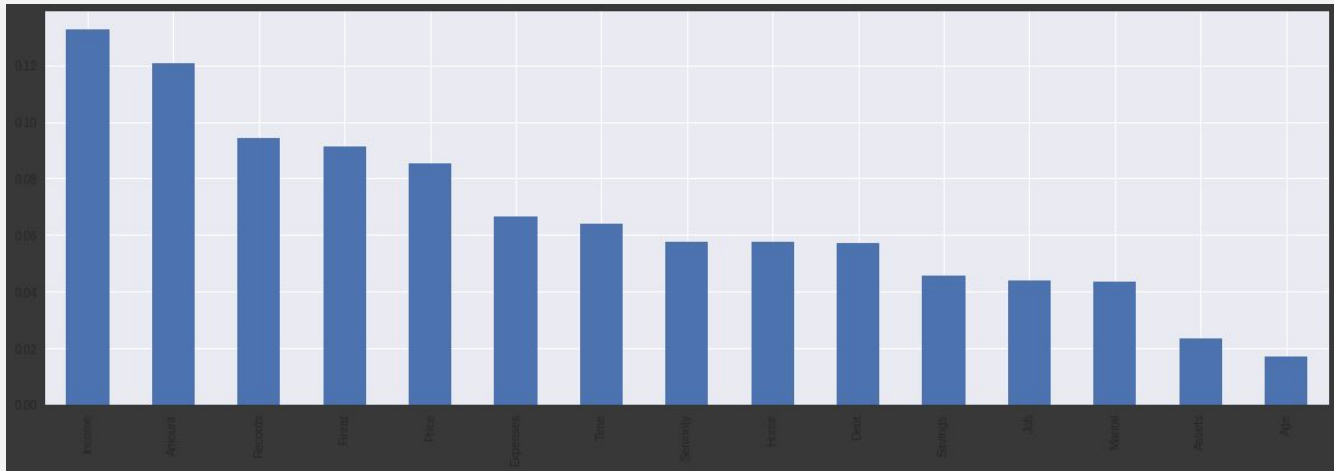
Some of the box plots are shown below:



**Studying impact of features on the target variable:**



**Feature importance:** We derived the importance of each variable and ranked them from the most to least important as shown below:



Also, we obtained the data for duplicate rows as shown below and removed them from our dataset and also checked if there were any duplicate columns. Now, our dataset contains around 4444 rows.

Duplicate rows	2
Duplicate rows (%)	< 0.1%

When we analysed the feature importances corresponding to the result variable, it was very difficult to distinguish between different variables as some of them shared almost equal feature importance. So, we

came up with another feature selection technique most suitable for our problem.

## Feature Selection:

We employed the technique of feature addition for feature selection in which we initially calculated roc-auc score considering all variables and stored it. Then, we started adding



one variable at a time from the most to least important and built an xgboost model at each round. If the increase between new roc-auc and previous roc-auc is more than a small threshold, we will keep the variable else we will discard it. This way, we filtered out the top 8 variables out of 13 which were most relevant to our problem domain. Those are:

```
['Income', 'Seniority', 'Records', 'Home', 'Debt', 'Savings', 'Job', 'Marital']
```

## STEP 5 (Model Building):

We selected the Random Forest Model as our baseline model. They are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean/average prediction of the individual trees.

```
# capture the 8 selected features
seed_val = 1000000000
np.random.seed(seed_val)

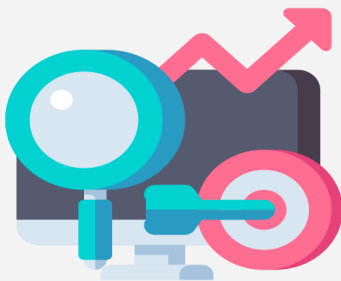
final_xgb = RandomForestClassifier()
final_xgb.fit(X_train[features_to_keep], y_train)

y_pred_test = final_xgb.predict_proba(X_test[features_to_keep])[:, 1]
auc_score_final = roc_auc_score(y_test, y_pred_test)
```

## STEP 6 (Performance Evaluation):

Test selected features ROC AUC=98.670000

## STEP 7 (Hyper-Parameter Tuning):



**The main parameters used by a Random Forest Classifier are:**  
 criterion: the function used to evaluate the quality of a split,  
 max\_depth: maximum number of levels allowed in each tree,  
 max\_features: maximum number of features considered when splitting a node, min\_samples\_leaf: minimum number of samples which can be stored in a tree leaf, min\_samples\_split: minimum number of samples necessary in a node to cause node splitting and  
 n\_estimators: number of trees in the ensemble.

```
from sklearn.model_selection import RandomizedSearchCV
n_estimators = [int(x) for x in np.linspace(start = 200, stop = 2000, num = 10)]
max_features = ['auto', 'sqrt']
max_depth = [int(x) for x in np.linspace(100, 500, num = 11)]
max_depth.append(None)
random_grid = {
    'n_estimators': n_estimators,
    'max_features': max_features,
    'max_depth': max_depth
}
from sklearn.ensemble import RandomForestClassifier
rfc=RandomForestClassifier()
rfc_model = RandomizedSearchCV(estimator = rfc, param_distributions = random_grid, n_iter = 100, cv = 3, verbose=2, random_state=42, n_jobs = -1)
```

Model ROC AUC improvement is 0.11026798

## GUI/DEMO

The screenshot shows a web browser window with the URL `bankingdata-analysis.herokuapp.com/predict`. The page has a dark blue background and is titled "Predict customer credit score". It contains a form with the following input fields:

- Your Age
- Job type (Fixed: 1, Part-time: 2, Freelance: 3, Other: 4)
- Job Seniority level (in years)
- Home ownership (Rented: 1, Owner: 2, Privilege: 3, Ignore: 4, Parents: 5, Oti)
- Marital status (Single: 1, Married: 2, Widow: 3, Separated: 4, Divorced: 5)
- Existence of earlier Records with bank (No: 1, Yes: 2)
- Per month expenses (in K)
- Monthly Income (in K)
- Number of Assets
- Debt Amount (in K)
- Amount requested for loan (in K)
- Price of good/purchase/any other purpose of loan (in K)
- Time period for requested loan (in months)

Below the input fields is a blue "Predict" button. To the right of the button, the following text is displayed:

According to our analysis considering the data provided by you, you seem to be a reliable customer to us with customer credit score: 76.97 percent

## CONCLUSION

We successfully completed the detailed analysis on how an actual credit scorecard predictor works and operates. We took inspiration from Lehman Brothers' case of bankruptcy and decided to present a step by step procedure to solve their major cause of their failure. In this report, we mentioned all our decisions and choices based on careful consideration of various factors relevant to our data. We reviewed various ways for exploratory data analysis, pre-processing, modelling as well as performance optimization methodologies. We also presented a user-friendly web application by which financial institutions can make two significant decisions: first, whether to grant credit to a new applicant, and second, how to deal with existing applicants, including whether to increase their credit limits.

# REFERENCES

[Lehman Brothers Bankruptcy Fraud](#)

[Credit Score - Facts, Myths & A Case Study](#)

[FICO Score](#)

[\(PDF\) An optimised credit scorecard to enhance cut-off score determination](#)

[CREDIT SCORING APPROACHES GUIDELINES](#)

[Credit Fraud || Dealing with Imbalanced Datasets](#)

[Credit scoring of real customers: A case study in Saderat bank of Iran](#)

[CASE STUDY: DEBIT AND CREDIT CARD FRAUD](#)

[Credit scoring - Case study in data analytics](#)

[Deploy to Production](#)

[Hyperparameter tuning for machine learning models](#)

[Credit scoring methods](#)

[Scoring Models of Bank Credit Policy Management](#)