**Objectives:**

The main objective of developing a Credit Card Approval Shiny App is to show the impact of different fields like Gender, Age, Income, Number of years employed etc on the approval for a Credit Card. This app have some static graphs(which include histograms, scatter plots, box plots etc) and some interactive plots that will help user to select the fields of interest.

The primary objective of this analysis is to implement the data mining techniques on a credit approval dataset. Risks can be identified while lending,data-based conclusions can  about probability of repayment can be derived and recommendations can be put forward.

**The aim of the analysis** is to look for impact the various fields have on our final outcome of Credit Approval.

**Look into the Data:**

The Credit Approval dataset consists of 690 rows , representing 690 individuals applying for a credit card, and 16 variables in total. The first 15 variables represent various attributes of the individual like fender, age, marital status, years employed etc. The 16th variable is the one of interest: credit approved(or just approved). It contains the outcome of the application, either positive(represented by “+”) meaning approved or negative (represented by “-“) meaning rejected. This dataset is a multi- variate dataset, having continuous, nominal and categorical data along with missing values.

**Data PreProcessing:**

Preprocessing of the data includes data cleaning, data integration, data transformation , data reduction, missing values imputation among other tasks. Below are some of the data transformations that were done to the Credit Approval dataset before we apply any EDA techniques.

1. The Credit Approval dataset contains categorical values that are transformed to binary values or factors of 1s and 0s. For eg., Approved field having values of + and – are changed to 1 and 0 respectively, 1 being the card is approved. Similarly, Gender having values ‘a’ changed to 1 representing male and ‘b’ changed to 0. Prior default and Employed both have categorical values ‘t’ and ‘f’ which are transformed to 1 and 0. 1 as binary value considered true/yes/pass and 0 represents false/no/fail.
2. Missing data: The missing values constitute to 5% of the entire dataset. And the missing values are represented by “?”. Converted all the missing values to NA first, and then imputed them(See below for more details)
3. Variable Names: Initially the fields were named from A1-A16 but with the help of some documentation available, there were renamed appropriately. However, All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data.
4. Data Types: Converted the ‘t’ and ‘f’ into factors.
5. ZipCode: The values for this field were mostly zeros or invalid. This field will not be considered for any analysis.
6. Number of Records: This dataset only has 690 observations, limiting us to come to a conclusion.

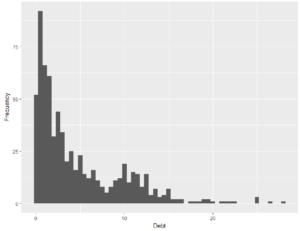
**Missing data Treatment:**

The missing values are found to exist in attributes Age, Gender,Marital Status, Bank Customer. Education level, Ethnicity, and Zip Code which we filled by NAs. Out of these, Age is a continuous variable. There are different methods to impute missing value, ranging from deleting the observations, deleting the attribute if of no importance, zero them out or plug the mean/median/mode value from all the values.

Here we imputed the values by using the median value for Numerical fields. For remaining attributes with categorical values, the missing values are imputed using the frequency count of the observations. The Class group with highest frequency was used.

**Exploratory Data Analysis:**

To start with, the distribution of 5 continuous variables Age, Debt, Credit Score, Income and Years employed was observed to get a sense of the nature of the dataset.

[](https://nycdsa-blog-files.s3.us-east-2.amazonaws.com/2018/07/pic4.png)

[Chart, histogram

Description automatically generated](https://nycdsa-blog-files.s3.us-east-2.amazonaws.com/2018/07/Rplot1.png)

These initial plots showed that all variables have distributions that are skewed to the right, indicating that the data is not well-distributed about the mean. In order to reduce the skew,log transformations were applied and then plotted again.

[Chart, histogram

Description automatically generated](https://nycdsa-blog-files.s3.us-east-2.amazonaws.com/2018/07/Rplot2.png)

[Chart, histogram

Description automatically generated](https://nycdsa-blog-files.s3.us-east-2.amazonaws.com/2018/07/Rplot3.png)

Below are the plots of the discrete variables that appear to influence whether a credit application is approved.

[Chart, bar chart

Description automatically generated](https://nycdsa-blog-files.s3.us-east-2.amazonaws.com/2018/07/pic8.png)

[Chart, bar chart

Description automatically generated](https://nycdsa-blog-files.s3.us-east-2.amazonaws.com/2018/07/pic9.png)

As expected, Prior Default and employment status appear to have the most significant effect on the approval. Persons with prior default are rejected more than 90% of the time, and those who not employed are rejected 70% of the time.

**Education**

Let’s see if the education level has any effect:

[Chart

Description automatically generated](https://nycdsa-blog-files.s3.us-east-2.amazonaws.com/2018/07/pic10.png)

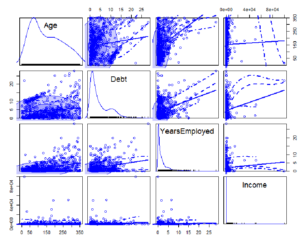
From the graph , we see that people with education level ‘x’ have an 85% chance of approval as-compared to ‘ff’ who are rejected 85% of the time.Among the continuous variables, Income and Credit Score seem to also have significant effect on the outcome of the credit application.

[Chart, histogram

Description automatically generated](https://nycdsa-blog-files.s3.us-east-2.amazonaws.com/2018/07/pic11.png)

As we see, a high credit score resulted in approval 90% of the time, and applicants with higher income have a higher than average approval rate.

Finally, I did a pairwise comparison of all the fields using a scatter plot you can see below:

[](https://nycdsa-blog-files.s3.us-east-2.amazonaws.com/2018/07/pic12.png)

These plots do seem to have a scaling problem. One reason for this could be the presence of outliers. The range of the values is high, causing the regression line to adjust for these outliers. For now, we will not be working on handling these. But from the plot, we can see that Years Employed has the highest linear correlation with the Approved field.

**Conclusion / Future Scope:**

From this initial analysis, we are able to conclude  that the most significant factors in determining the outcome of a credit application are Employment, Income, Credit Score and Prior Default.

Based on these insights, we can work on building some predictive models. They can be used by analysts in financial sector and be incorporated to automate the credit approval process. These results can also serve as a source of information for the consumers.

Modern credit analyses employ many additional variables like the criminal records of applicants, their health information, net balance between monthly income and expenses. A dataset with these variables could be acquired. It’s also possible to add complementary variables to the dataset. This will make the credit simulations more  , similar to what is done by the banks before a credit is approved.

**My Day to Day activities:**

the most technical aspect of my job is collecting the data itself. I am working together with web developers to optimize data collection

My job also included performing data analysis using Excel, and Python and creating data visualization reports using Tableau with relevant charts and graphics.

In short, as a part of this project, I worked with different SQL queries which involves using select statements, aggregate functions, joins, and subquery. I’ve used matplot lib with Seaborn to make statistical plots more attractive. I have also used Data visualization tools like Power BI and Tableau to build reports and dashboards.

As an analyst, I spent a significant amount of time producing and maintaining both internal and client-facing reports. Those reports give management insights about new trends on the horizon as well as areas the company may need to improve upon.

The other significant part of my job involves creating weekly, monthly, and quarterly reports for Internal Audit issues and Regulatory issues and dashboards for senior management

Looking forward, I am seeking opportunities that will give me a chance to impact the business strategy by recommending the best choices based on data.