**Case 1: Analysis on a Lending Club Platform**

Nikita Goel

**Table of Contents**

[**Dataset Description** 3](#_Toc100602733)

[**Issues with the dataset** 3](#_Toc100602734)

[**Visualizations and observations** 3](#_Toc100602735)

[**Feature set selection** 8](#_Toc100602736)

[**Data cleansing and Model creation** 8](#_Toc100602737)

[**Test results** 9](#_Toc100602738)

[**Proposed enhancements** 11](#_Toc100602739)

# **Dataset Description**

This dataset is providing information about the loans given to different customers including their financial background and specific loan information about each customer. The dataset is well comprehensive with 55 variables and 10000 data points. The Lending Club platform disperse loans to customers depending on their financial stability. Some customers can get high interest rates, and some can get lower interest rates depending on their pay back surety. This dataset contains the outcomes of the loans approved, where some data points can also contain information about the users who refused to take the loan or did not get the loan.

# **Issues with the dataset**

* A lot of empty of NA cells are contained in the data which require additional pre-processing for each variable.
* Some columns contains a lot of records with the value 0 (like number\_collections\_last\_12m, current\_accounts\_delinq, paid\_late\_fees etc) which might not provide any meaningful information.
* Some columns do not contain any data and are blank (like verification\_income\_joint). These columns might not provide any meaningful insight and could not be used in algorithms since they do not cater the complete population.

# **Visualizations and observations**

Here is the link to the visualizations created while analyzing this dataset (<https://public.tableau.com/app/profile/nikita.goel/viz/casestudy1_16497261047750/visualization1_1>). The following are the topics for visualizing the dataset:

1. **Is there any relationship between average annual income and loan interest rates?** (Figure 1)

Observation – It is a very interesting visualization showing a considerable dependence of annual income to determine the interest rate. Figure 1 shows that the highest average annual income employees receive a much lower interest rate and vice-versa. As the income decreases, the rate of interest increases and at a certain minimum value of the income, the trend starts increasing a bit but still for lower- and middle-income level, the rate of interests is higher.

1. **Which state to give attention regarding different loan statuses** (Figure 2, 3, 4)

Observation – This visualization will require a choropleth chart with an interactive dashboard having a list to choose from different loan status types. With each status type changing, the colors of different parts of the map change to depict the density of loans in that area. Interesting results were obtained while observing the behavior of the dataset in Tableau (Figure 2, 3, 4). California has almost double the loans than New York (Figure 2). In this, California is the state with most loans, but when it comes to Late repayment within 15-30 days, New York becomes has the highest density of that loan status type (Figure 3). For the status Late repayment within 31-120 days, these two states are equivalent in number (Figure 4).

1. **Visualizing the interest rate per state** (Figure 5)

Observations - Different colors in the choropleth map showed the different ranges of the rate of interest. Dark green indicated the lowest interest rate, yellow the average interest rate and red the highest interest rate. Visualizing the interest rate per each state gave a meaningful insight where North Dakota and Hawaii indicated the highest interest rates (marked in red) as compared to all the other states. Alaska and Nebraska showed the lowest rates of interest marked in dark green, as compared to all the other states.

1. **Plotting the interest rate variation based on the loan amount** (Figure 6)

Observations – Based on the loan amount, an analysis was made that whether the interest rate can vary. It is quite true that different loan amounts have different interest rates. Another observation is that after every short interval of increasing loan amount, the interest rates spike which means that they increase to a much larger extent and then drop. Similar trend can be observed with the downfall of the plot as well.

1. **Distribution of applications over different types of loan** (Figure 7)

Observations – The plot showed the distribution of different types of loan taken by either individual or joint applications. The percentage of joint loans taken as compared to the individual loans is much lower for each type. Debt consolidation loans are taken by the maximum people. Loans on renewable energy are the least.

1. Interest Rate vs Annual Income (Figure 1)

Graphical user interface, chart

Description automatically generated

Figure 1: Interest rates based on annual income

1. Interactive dashboard showing different loan statuses based on location (Figure 2, 3, 4)

Graphical user interface, application, map

Description automatically generated

Figure 2: For all loan statuses

Graphical user interface, application, map

Description automatically generated

Figure 3: For Late (16-30 days) loan status

Graphical user interface, application, map

Description automatically generated

Figure 4: For Late (31-120 days) loan status

1. Visualizing the interest rate per state (Figure 5)

Graphical user interface, application

Description automatically generated

Figure 5: Visualizing in Tableau for average rate of interest per state

1. Plotting the interest rate variation based on the loan amount (Figure 6)

Graphical user interface, chart

Description automatically generated

Figure 6: Visualizing variation in interest rate with increasing loan amounts in Tableau

1. Distribution of applications over different types of loan (Figure 7)

Chart

Description automatically generated

Figure 7: Visualizing in Tableau the application types for different types of loan

# **Feature set selection**

The following variables were included in the feature set:

* Annual Income
* Total credit limit
* Total debit limit
* Loan Amount
* Term
* Installment
* Interest Rate

# **Data cleansing and Model creation**

The assumptions made while building this model was the feature set variables which were picked while studying the dataset. Since I used two algorithms Regression model using Decision Trees and Linear Regression implemented by Spark’s Machine learning library, I wanted the input file to be of the format LibSVM. For cleaning the data, the following process was followed:

* The feature set were chosen from the dataset after studying the dataset and analyzing the columns that can assist in prediction.
* The dependent variable was put in a text file in the beginning and in the last column, and the column number was appended along with the column value separated by space to convert the dataset into LibSVM format.
* After cleaning dataset, it was loaded into the data bricks python notebook.

Text, email

Description automatically generated

Figure 8: Regression model using Decision Trees code

The code was created in data bricks where I created cluster and loaded the LibSVM dataset into the cluster. I fitted the model into the regression model using decision trees algorithm and executed the code.

Two algorithms used were as follows:

1. Decision tree regression (Code in Figure 8)
2. Linear regression (Code in Figure 9)

Graphical user interface, text

Description automatically generated

Figure 9: Linear regression model

# **Test results**

1. Test results of the Decision trees regression model (Figure 10) is shown below:
2. Test results of the linear regression model (Figure 11, 12) are shown below:

Graphical user interface, text, email

Description automatically generated

Figure 10: Output of the decision tree regression

Graphical user interface, text, email

Description automatically generated

Figure 11: Part 1 of the linear regression output

Graphical user interface, text, application, email

Description automatically generated

Figure 12: Part 2 of the linear regression output

# **Proposed enhancements**

* The charts can be built in Python Bokeh and more customizations can be built beyond the scope of the Tableau tool.
* More sophisticated machine learning algorithms can be used to predict the interest rate. These algorithms can be compared with each other and more accurate one can be implemented.
* More comprehensive visualizations can be made from this diverse dataset including other variables and making meaningful insights.