Data Analytics for Lending Club

A project by Vishal Satam

Summary

This project is a comprehensive study on the dataset provided by Lending Club. Lending Club is a peer to peer money lending website which invites investors to invest money into their company and these funds are utilized by lending a portion of it to different types of borrowers. Lending Club makes its money from the interest that is paid by the borrowers over time and pays a portion of this money back to the investors. There is a considerable difference between the amount of interest charged to the borrowers and the amount of money paid back to the investors. Lending Club maintains this difference as contingency against defaulters. Exploratory, Prescriptive and Predictive analysis has been performed on the available datasets which aims at profiling the lending behavior by utilizing the design patterns of the Map Reduce framework. The main focus of this analysis has been to determine an investing strategies and to summarize and create a profile for lending club which could seem attractive to investors. The variation of interest rates has been analyzed and an effective model has been developed to predict interest rates using linear regression. A correlation of the US inflation rates and the interest rates has been performed to indicate if the inflation changes affect the business of Lending Club. Feature determination for predicting loan default has been carried out by performing exploratory analysis on the dataset to determine the best features that could increase the accuracy of predicting whether a loan will default using Multiclass Decision Forest Classification algorithm. Sentiment analysis was performed on the descriptions provided by the borrower as a detailed view of the purpose of the loan and a polarity score was assigned to each loan to explore a correlation exists with the loan status, i.e Default or Fully Paid. Different types of charts and Graphs have been developed that help visualize the analysis performed. The main challenged faced was the way the member ID and loan IDs are assigned in the dataset. Since every loan is unique, borrower profiles in terms of multiple loans borrowed based on any unique ID could not be formed. Also, as is with most data analysis, valuable data extraction from the dataset was also a major hurdle and increased the analysis time. The available data was cleaned and transformed according to the requirements of our analysis. Intelligent techniques and workarounds were developed to clean the data and handle missing information in order to reduce time spent in debugging parsing issues.

Datasets Used

Lending Club dataset

https://www.lendingclub.com/info/download-data.action

https://www.kaggle.com/wendykan/lending-club-loan-data

https://resources.lendingclub.com/LCDataDictionary.xlsx

Rejected Loans

https://www.lendingclub.com/info/download-data.action

US Inflation 2006-2017

http://www.usinflationcalculator.com/inflation/current-inflation-rates/

List of positive and negative words

https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar

Preparatory Data Cleaning

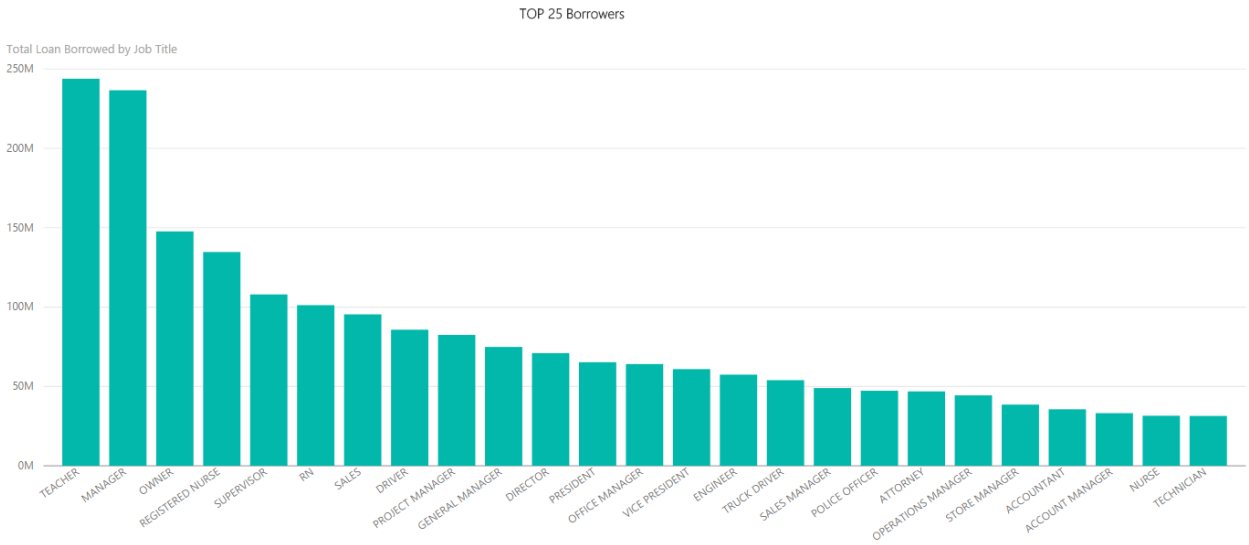
1. Developed a UtilityFunctions class to develop a customized effective parser that suits the main dataset of Lending Club.
2. The values in all the date fields were re-formatted by writing a generic function which was added in the UtilityFunctions class. This function transforms all the dates into MMM-YYYY format from the inconsistently occurring MMM-YY and/or YY-MMM format to prevent future ambiguities as well as issues in BI tools.
3. Separated the dataset into 2 parts, one that contains only the descriptions and the ID and the other with the remaining fields to avoid parsing issues and to increase performance as most of our analysis does not require the loan descriptions. Performance gain is observed because most of our analysis is performed on the main dataset which doesn't contain the descriptions.
4. Purpose has been cleaned to include the correct types of loans that Lending Club offers.
5. Term has been cleaned to remove "months". This is redundant and adds space overhead.
6. We are not doing anything for missing values because these will be handled when we parse and analyze the individual fields.
7. Loan title has been cleaned to remove unnecessary double quotations and commas to reduce parsing issues.
8. Removed URL from the dataset as we don't require this.

Tools Used

1. NetBeans
2. Hadoop Map Reduce Architecture
3. Power BI
4. Microsoft Azure Machine Learning Studio

Analysis Performed and Results

1. **Top 25 Borrowers**



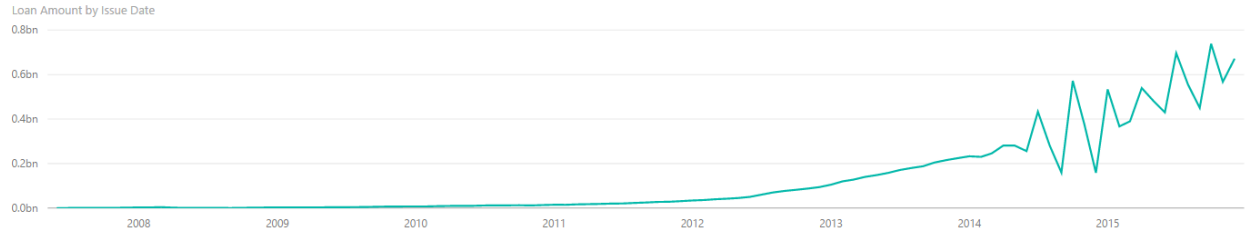
Purpose : This analysis was performed to determine what emp_titles constitute the top 25 borrowers. The value for this is calculated using the sum of the funded amount for the loans. This analysis helps us understand the borrower profiles based on the total amount of loan borrowed.

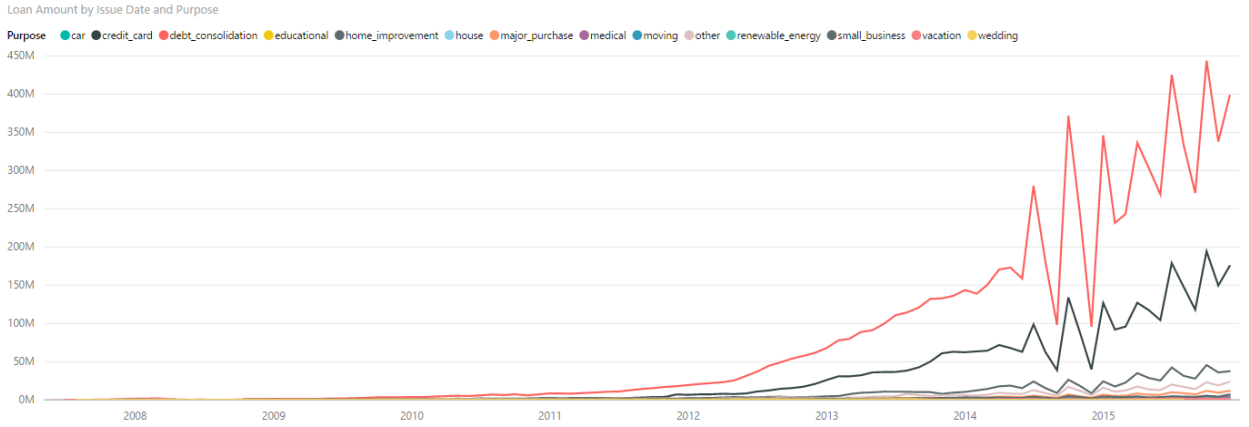
Result : The analysis conclusively determines that Teachers and Managers have the highest borrowed loans and Lending club should focus it's marketing campaigns towards people from these professions. This might lead to increase in sales of loans.

Map Reduce Design patterns : Filtering pattern - Top 25 analysis, Job Chaining to sort borrower amount in descending order and filter out top 25 borrowers. Secondary Sorting using Composite Key for Borrower and the sum of the amounts of loan.

2. Exploratory analysis on the total loan amount issued by Date

Purpose : This analysis is performed to gain an insight into how the Lending Club is performing overall. A good indicator of this more than the overall profit made is the amount of loans that Lending Club is issuing. This analysis will help us understand the sales volume and investor confidence.

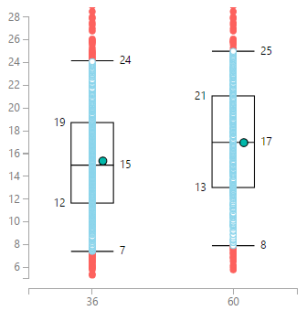




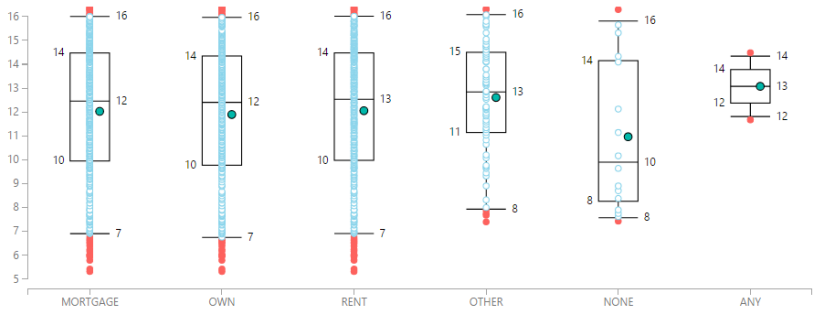
Results : The analysis performed above confirms the understanding that Lending Club's business is growing steadily since inception. We can also see that since 2012, there is an exponential increase in the amount of loans issued by Lending Club. These are positive indicators which point to increasing investor confidence and increasing volume loans being sanctioned.

Map Reduce Design Patterns : Combiner Optimization using a single implementation for Reducer and Combiner class.

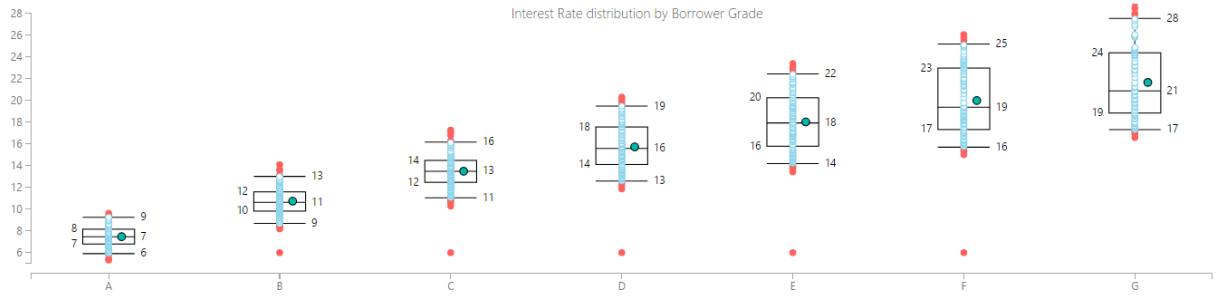
3. Exploratory Analysis on the average rate of interest grouped by various factors

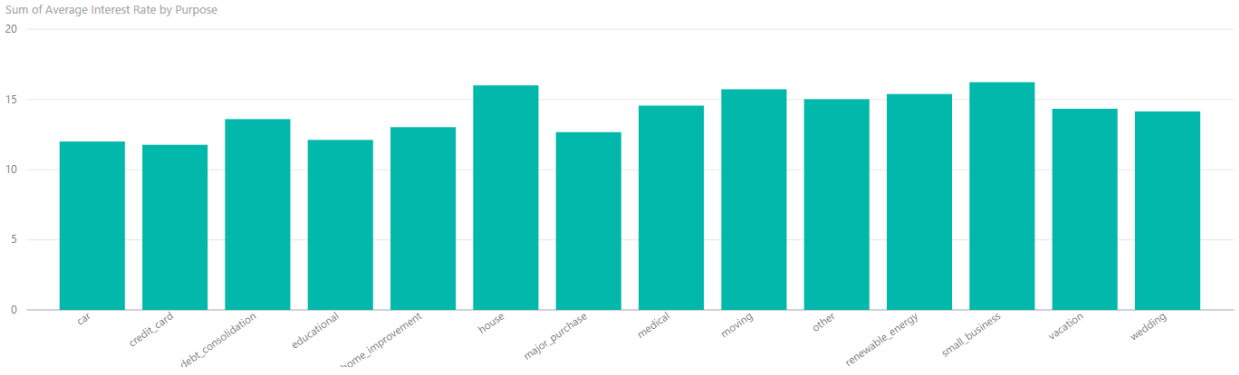


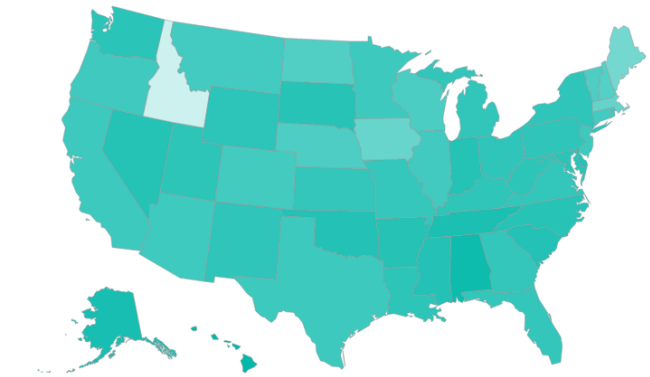
Interest Rate distribution by Term

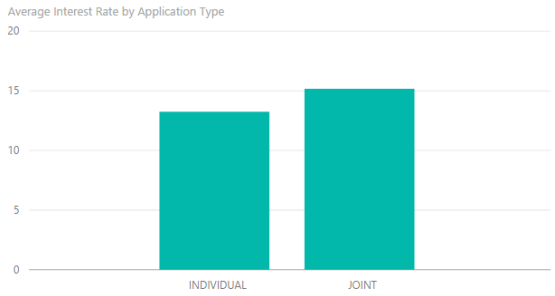


Interest Rate Distribution by Home Ownership







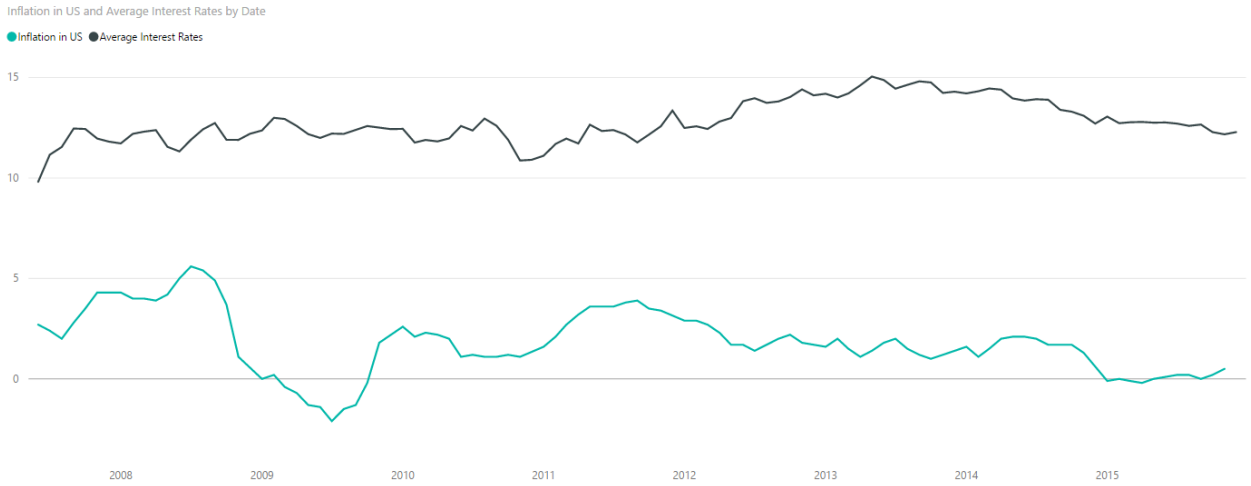


Purpose : Interest Rate is the most important attribute for Lending Club's business because this is what drives their income and helps to attract borrowers. This exploratory analysis performs summarization and computes the average of interest rates grouped by various factors such as term, home ownership, state, borrower grade, purpose and application type. These summarizations were performed to determine features that could adequately predict or affect the interest rate of a loan.
For this analysis, the different features were emitted as key to be grouped and the interest rate was emitted as the value.

Result : From the analysis conducted, we can clearly see a relationship between the borrower grade and the interest rate.
The term of the loan has some effect on the rate of interest as we can see that the longer the term, higher the interest. We can also see that if the application type is Individual, the interest rate is lower as compared to the Joint applications. Although these factors seem to affect interest rates, the difference between the interest rate values is not considerable enough to be taken as features for regression.
We can see from the whisker plots and the bar graph that Home ownership, State and Purpose do not affect the interest rate at all because the variation of the interest rates per feature type is not profound and cannot be considered as contributing features for interest rate.

Map Reduce Design Patterns : Numerical Summarization pattern for calculating Average of Interest Rates

4. Correlation of Inflation rate in US with the Interest Rate



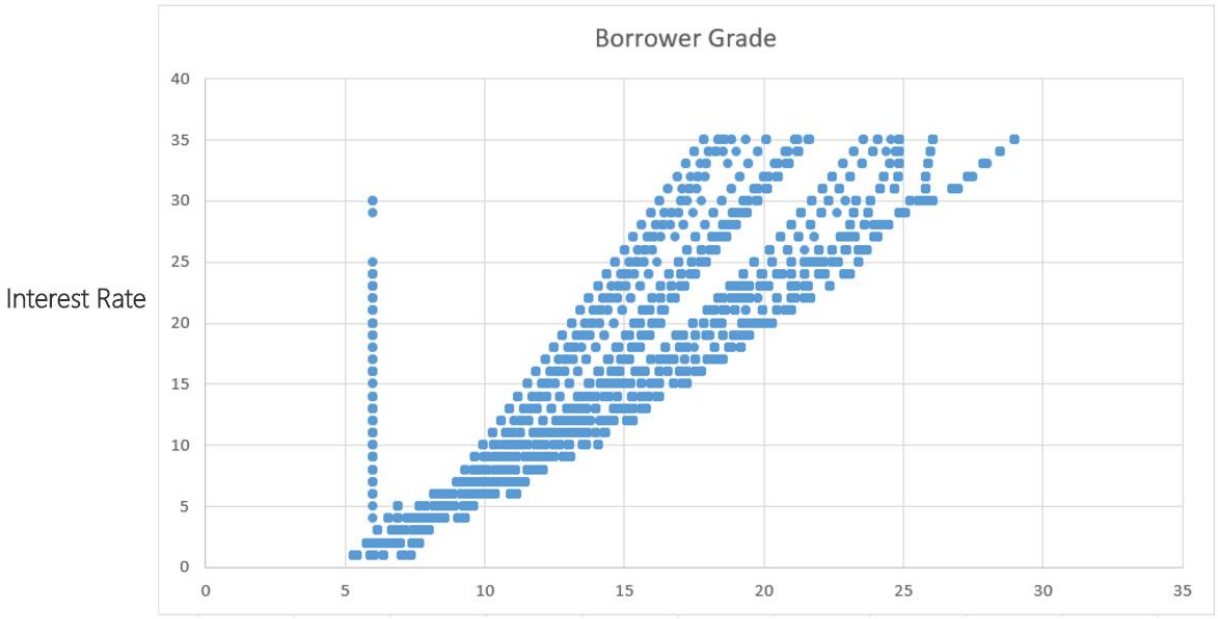
Purpose : This analysis was carried out to explore Correation between the Inflation rate in the US and the Average Interest rate being offered at Lending Club. The general rule of economics is that as the inflation rate rises, the interest rates drop and vice versa. This analysis would give us another feature, external to our primary dataset of Lending Club which might be important in predicting the interest rate. A map reduce job was created for determining the average rate of interest grouped by the issue date because it was a foreign key to be used to perform a map-side join with the US Inflation rates data. A map-side join was performed to determine correlation between the two datasets. Since the size of the inflation dataset was small, the Distributed Cache was utilized to improve performance by choosing a map-side join over a reduce-side join.

Result : No correlation was found between the inflation rate and the average interest rate when mapped by the issue_date of the loan.
We can observe that between June 2008 and June 2009, the inflation rate in US dropped by a huge margin, in fact it became negative. But, contrary to our expectation, the average interest rate for Lending Club loans remained unaffected by this dramatic drop in inflation.
From this, we can conclude that, since Lending Club is an autonomous peer to peer lending institution, which directly conducts business independently of the government with investors and borrowers, the interest rates do not fluctuate with fluctuations in the Inflation.

Map Reduce Design Patterns : Numerical Summarization pattern for calculating average of interest rates. Combiner optimization with single implementation of Reducer used for both Reducer/Combiner. A Writable class was created to facilitate the single Reducer/Combiner implementation.
Job chaining used to pass the output of the Average computation map reduce job as the input to the MR job that performs the Map side Join Pattern with Distributed Cache.

5. Predicting Interest Rate using Linear Regression

Purpose : This analysis is to provide a prediction model based on which we can predict the interest rate of a loan application. We have been building up this predictive analysis so far by performing exploratory analysis to find out the appropriate features which can be used to predict the rate of interest of an application. We have seen that the Borrower grade is the most important feature that can determine the Interest rate associated with the loan application.
The dataset has 2 grade related features, i.e the Borrower Grade (grade), and the Borrower Sub-Grade(sub_grade).



Data Preparation : We have given numerical values to the features in order to represent them accurately in increasing order and primarily so that they can be used as the independent variable in the linear regression model. Grade was assigned as multiples of 5 starting with A as 0, B as 5 till G as 30 and the sub-grade was assigned increasing numerical values from 1-5. And the total grade is the summation of the grade and it's corresponding sub-grade. So for example the grade-subgrade combination A3 is transformed as 3 (A=0, +3) whereas the grade-subgrade combination D2 becomes 17 (D=15, +2).

We can see in the above scatter plot that the interest rate is linearly correlated with the borrower grade.

Model derivation using Linear Regression

Hence, this is the perfect candidate for performing linear regression with borrower grade as the independent variable and the interest rate as the dependent variable.

Job chaining is used to create 3 jobs to compute the linear regression model.

- Job 1 : A map reduce job was created to perform this data transformation and also utilize the Partitioning Pattern to divide the dataset into train and test datasets. SRS(Simple Random Sampling) Filtering pattern has been used by applying a filter value of 90% to split the data into a train and test sub datasets using a Partitioner. Random sampling is used for this because if we

obtained a random sample of the data as train data, then this train sub dataset would represent the entire dataset adequately

- Job 2: In this job, we iterate over the train dataset which was produced from the previous job and calculate the mean value of the interest rate as well as the mean value of the borrower grade using combiner optimization by using a single implementation for Reducer/Combiner. The output of this job is put into the Distributed Cache because it only contains the mean value of both, interest and mean along with the number of elements. The mean is computed so that it can be used in the formulas used to derive the model.
- Job 3 : This job reads the output of the previous job, i.e the mean of the interest and the borrower grade and iterates over the train dataset to derive the model based on the following.

$$R = \frac{\sum^{n}(x-\bar{x})(y-\bar{y})}{\sqrt{\sum^{n}(x-\bar{x})^2(y-\bar{y})^2}} \qquad SD_y \ \sigma_y = \sqrt{\frac{\sum^{n}(y-\bar{y})^2}{(n-1)}} \qquad SD_x \ \sigma_x = \sqrt{\frac{\sum^{n}(x-\bar{x})^2}{(n-1)}}$$

$$b = \frac{R\sigma_y}{\sigma_x} \qquad a = \bar{y} - b\bar{x}$$

Model

$$\boldsymbol{y = a + bx}$$

In the above equations, $\bar{x}$ and $\bar{y}$ denotes the mean and σ_y and σ_x denote the Standard deviation of x and y respectively.

Results : The following model were obtained as a result of the above computation.

R	0.977307
R^2	0.95513
Interest Mean	13.24614
BorrowerGrade Mean	11.95847
Number of Elements	798645
Standard Deviation Interest Rate	4.381168
Standard Deviation Borrower	6.491925
Intercept	5.359254
Model	**Interest = 5.359253664637565 + (0.6595225194868153 * BorrowerGrade)**

The above model can be used to predict the value of the interest rate based on a given value of borrower grade as (A3 or F5 or B4 … so on)

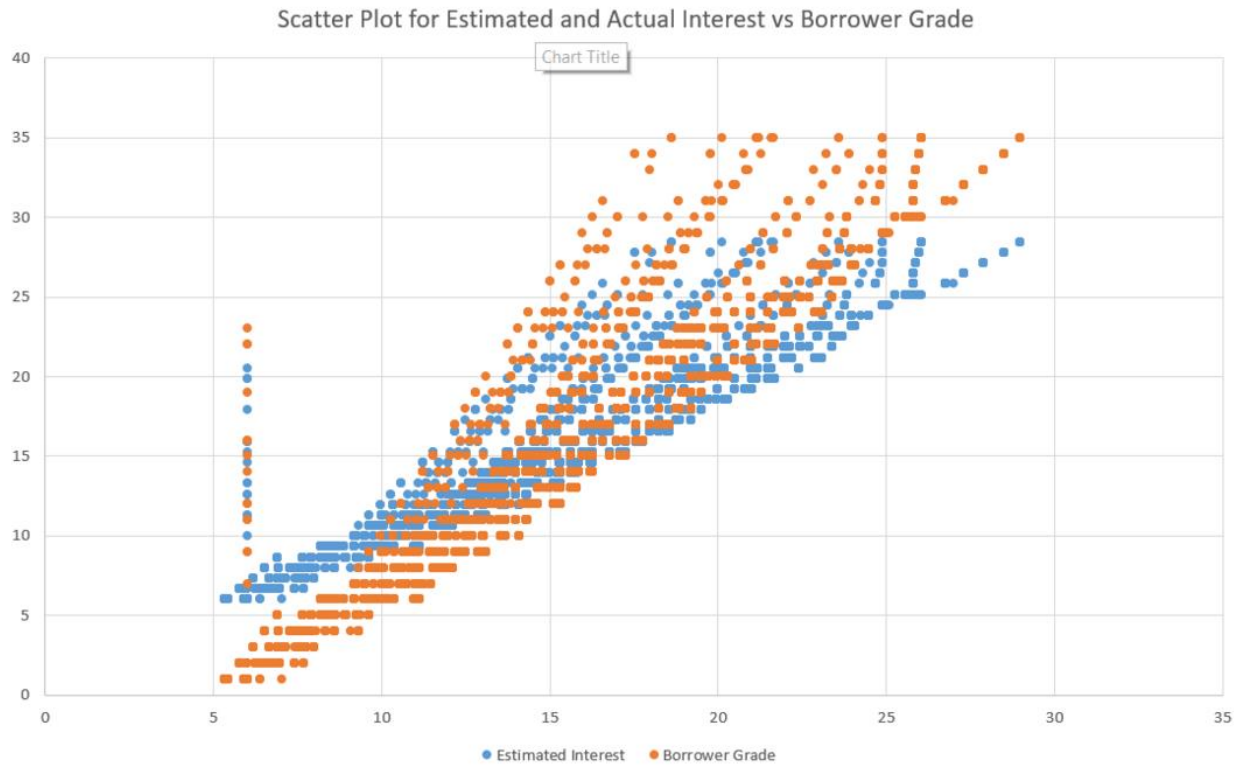
For the purpose of scoring the model, another MR job was created. This job uses the following formula to calculate the Standard error of the model that we derived in the previous step. Combiner optimization could not be used in this step because we required the entire sum of values before computing the standard error.

$$Standard\ Error = \sqrt{\frac{\sum^{n}(\hat{y}-y)^2}{(n-2)}}$$

Where $\hat{y}$ is the estimated value, y is the actual value and n is the number of elements on which this model was scored.

The following standard error values were calculated in the above MR job.

Std Error	**0.926981**
Std Error Squared	**0.859293**



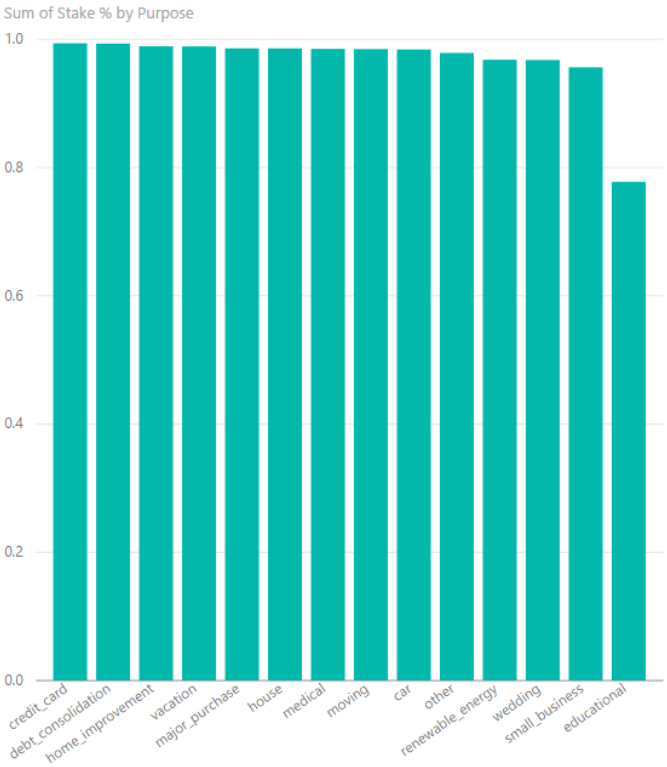
The above scatter plot has been plotted based on a Map only MR job that was run to calculate the estimated values of the interest rate based on the borrower grades present in the train dataset that was prepared during the split phase of the data preparation. The coefficients a and b that were computed in the previous model derivation step have been provided as inputs values in the context.

Map Reduce Design Patterns : Job Chaining, Numerical Summarization (Standard Deviation and Average), Data Organization Pattern (Partitioning), Filtering pattern used to Split data, Combiner Optimization using a single implementation for Reducer/Combiner and created a custom Writable class for outputting values from Mapper and Combiner.

6. **Improving Investor Strategy - Prescriptive Analysis**

Purpose : This prescriptive analysis is performed to analyze the current status of the different aspects of the Lending Club dataset that an Investor might be informed of to make better decisions in order to improve investment strategy to maximize profit and minimize default.

a. For this analysis, we first check the stake ratio of the investors to understand which type of loans are ignored or favored by the investors. We then perform exploratory analysis on the data to come up with suggestions to improve investor strategy. A MR job was configured to calculate the difference in the funded amount for a loan and the amount funded by investors to understand the percentage of the investors stake in a loan. The values of these stakes for all loans of a type of loan are averaged and the below bar chart is produced for visualization.

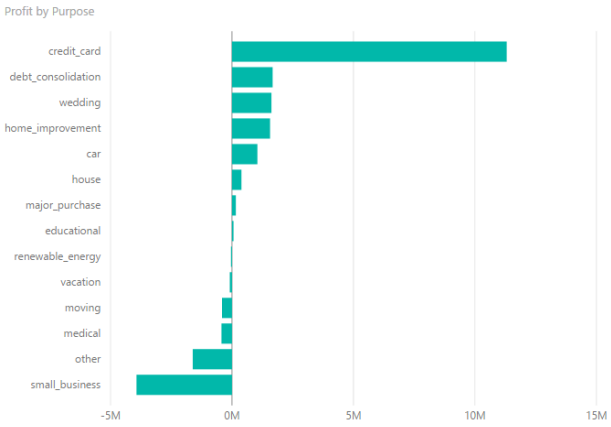


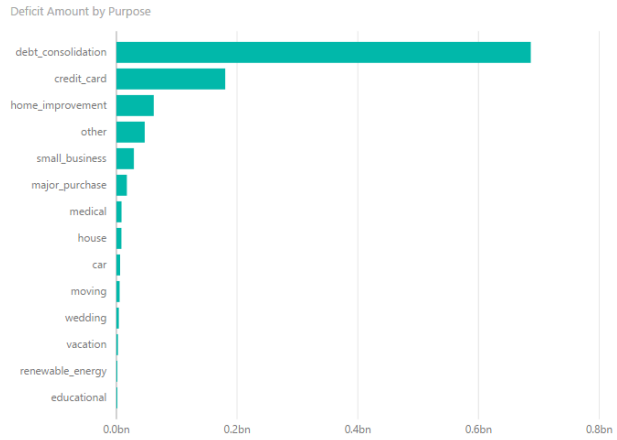
Result : From the above investor stake chart, we can see that the investors are mostly diversified and have a high percentage of stakes in every type of loan. The most intersting fact that we can derive from this chart is that there is less interest among investors for educational loans. We will attempt to understand more on the condition of Educational loans from a investor viewpoint.

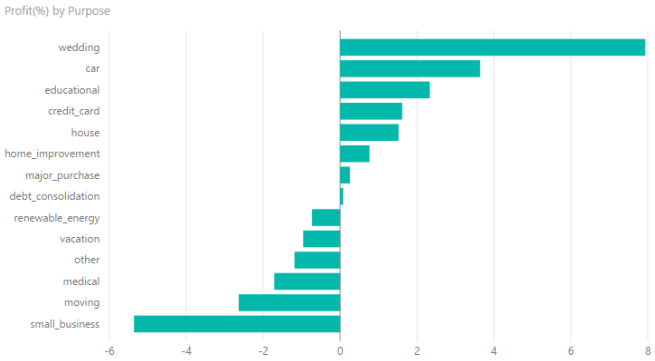
Map Reduce Design Patterns : Combiner Optimization used with single implementation for Reducer/Combiner while determining Stake Ratio.

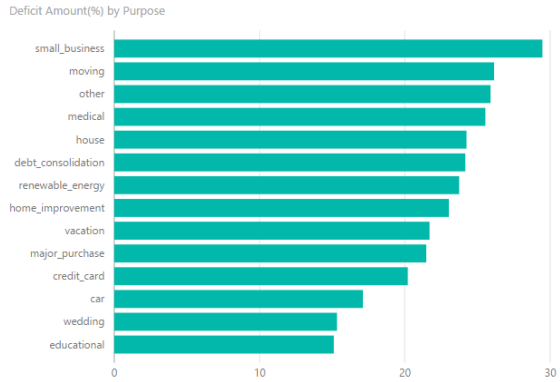
b. We then move on to analyzing the profit and the deficit amount between the expected income and the actual income. These values can give us a better indication of the current state of Lending Club's portfolio. We calculate 4 major parameters for this analysis. The profit percentage and amounts per purpose and the Deficit amount and the deficit amount percentage. The Profit is calculated as the difference of the total amount of money received

from the borrowers and the amount of money funded by Lending Club. The received money includes the interest that we get from the borrowers. The deficit amount is the difference between the projected income (sum of the amount of money which that type of loan could have fetched) and the actual income (sum of the amount money actually received back from the borrowers).









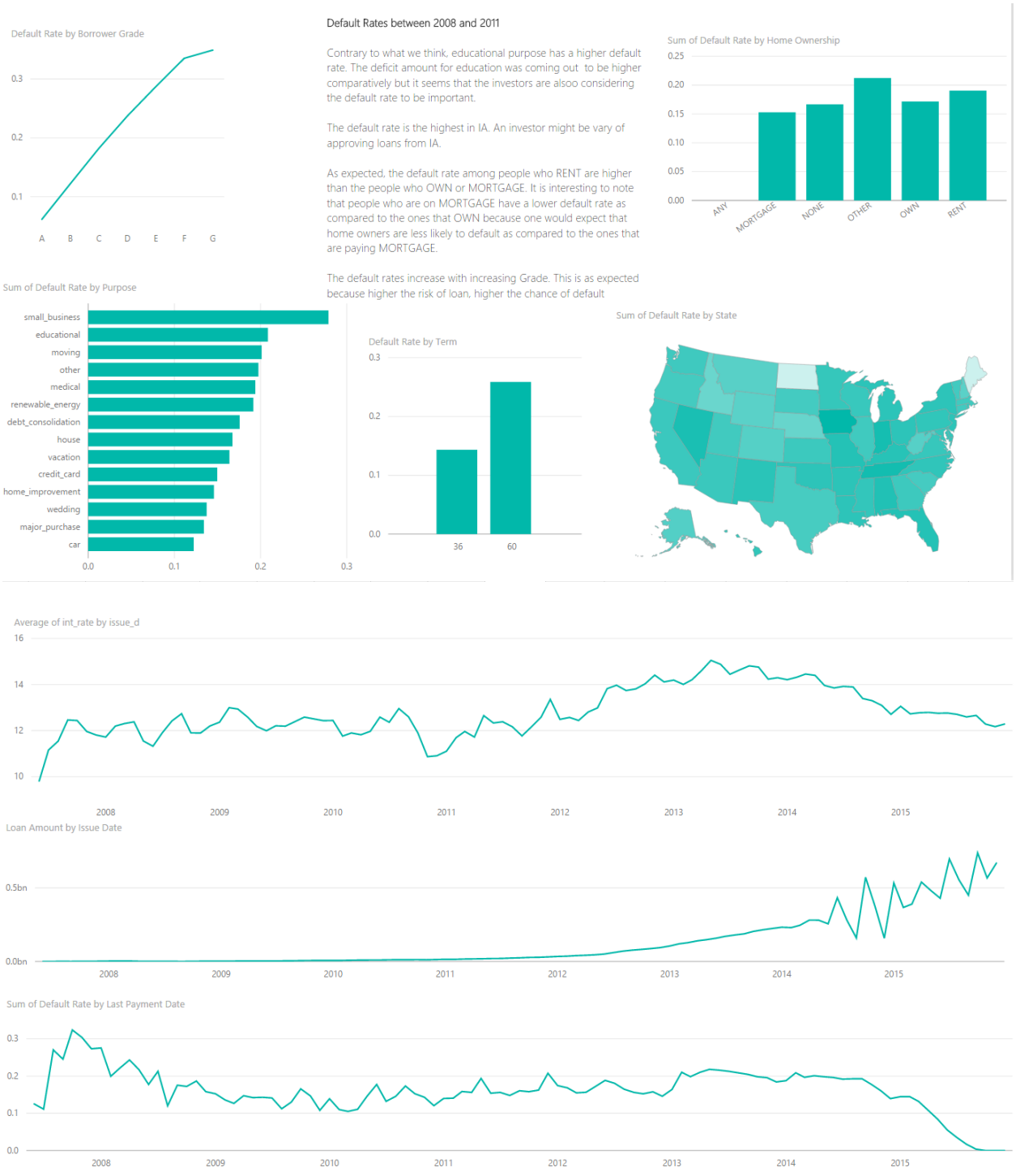
Result : When we see this analysis, it is quite evident that Educational Loans don't have large sums of money coming in as profit as compared to credit cards or debt_consolidation but the deficit amount, i.e the difference between the money that could have been made vs the money actually made is the least in Educational Loans. This means that the risk of default of the difference in the actual income and projected income is lesser as compared to say small businesses.

Educational loans can actually be looked at as a viable option in future to increase profits and have better estimations and results in terms of projected income. This way the investors get to meet their quarterly goals for achieving profit.

Another observation that we can make is that currently, the investors seem to be taking an aggressive stance by investing in credit cards. The deficit percentage is quite low. Also, credit cards bring in the highest profit as compared to any other type of loan. This might be a good thing, but according to the deficit amount, credit cards and debt consolidation are way higher. The investors might want to consider moving away from these types of loans by taking a moderate stance and look at investing in loans that have lower deficit amount thereby having a lower chance of defaulting.

Map Reduce Design Patterns : Numerical Summarization patterns used to compute values for profit, profit %, deficit and deficit %. Custom Writable created for writing funded amount (investment), amount received (income) and projected profit.

7. Loan Defaults Exploratory Analysis



Purpose : This analysis is performed to understand characteristics of loans that default. A good indicator to understand this is the default rate of the loans in our dataset based on different factors such as Borrower Grade, Home Ownership, Purpose, Term, State. We are trying to determine which features affect the default rate by plotting the variation by the feature.

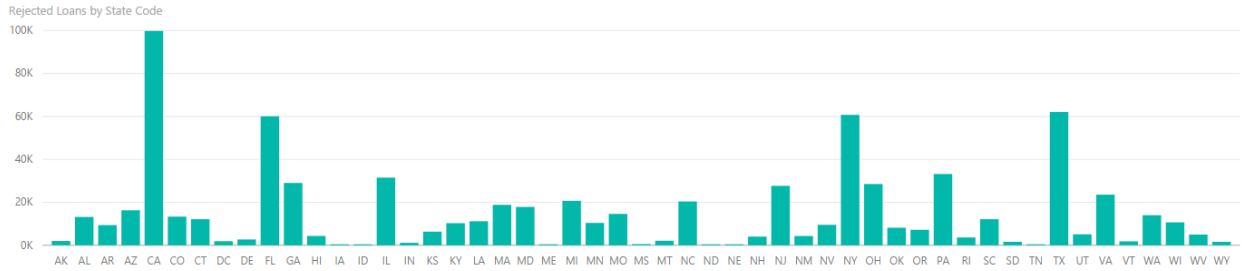
Results : As we can see in the above charts, the default rate has a direct linear relation with the borrower grade. We have also seen this pattern while performing linear regression for predicting interest rates. This outcome is as expected because as the grade increases from A to G, the risk associated with the borrower profile increases and hence the probability of default tends to rise.

The default rate has remained unaffected and roughly the same since the beginning of 2008 till date. This is an indicator that Lending Club is doing something wrong in terms of selecting loans. According to this statistic, we can conclude that we must try to come up with a predictive tool to predictLoan Defaults.

The default rate also seems to vary by the home ownership. As expected, the borrowers who RENT are more likely to default as compared to borrowers who are home owners or are on mortgage. It is interesting to note that the borrowers on mortgage have a lower default rate as compared to borrowers that own homes.
Loan default rates change considerably dramatically by purpose and there seems to be a linear correlation between purpose and the default rate. It is interesting to note again that Education has higher default rates. It is possible that the investors are being wary of investing in loans that come under that purpose. **As a result of this, we are left with a confusion as to whether a higher deficit amount and lower profit is a better option or should a higher loan default ratio matter.**
We can observe that the loan default rate is higher for loans that have a higher term. This is expected because people tend to default more when they have a larger overall amount to pay back. And there could also be external factors which affect the loan re-payment.
It is very straightforward to notice that the state in which the loan was issued does not have much effect on the loan and there is little variance between the loan default rates of different states.

Map Reduce Design Patterns : Numerical Summarization

8. **Correlation of Rejected Loans and Default Rates per State**

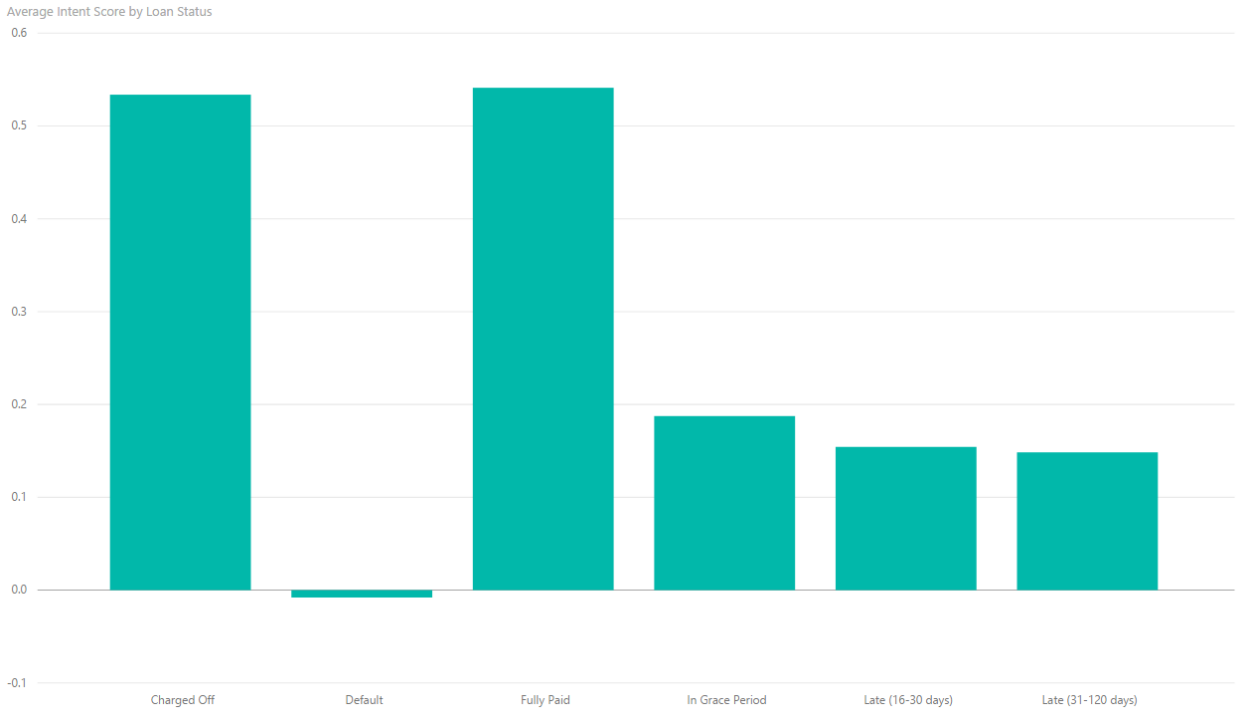


Purpose : This analysis is to determine if there is a correlation between the loan default rate and the number of loan rejections per state. An inner join was performed on the Rejects dataset and the default rate by state that we obtained during the previous analysis.

Results : The hypothesis was that the number of rejections would rise if the default rate in a state was high but this analysis revealed no concrete evidence to support this claim.

Map Reduce Design Patterns : Reduce side join pattern.

9. Intent Analysis of Loan Descriptions provided by Borrower



Purpose : The borrower usually provides a detailed description of the purpose with the loan application. This can be considered as an important feature in order to form a profile of the borrower to analyze and rate the borrowers intent. To rate a borrower's intent, intent analysis is performed on the loan description. A score of +1 is assigned for every positive word and a score of -1 is assigned to every negative intent word. The overall intent score is calculated by adding the scores of the individual words. A map reduce job is run to join together the loan descriptions dataset and the main dataset to have the Loan Status and the Descriptions in one file. The ID is used as the foreign key to perform the inner join over these two datasets.

We use the Bloom Filter MR filtering pattern to check for matching intent words in the positive and negative intent words list. This was done to improve performance.

Result : The intent score has been calculated during the Map phase and the average of all the individual scores has been computed in the Reduce phase. The key Loan Status and value of intent score per loan was emitted by the Mapper and used in the reducer to compute averages of the intent score.

The conclusion of this analysis is that the loans that are in the Fully Paid status have a higher intent score on an average as compared to the loans that have Defaulted. This can also be used as a feature or

parameter in predicting or to at least caution to Lending Club as to exercise more scrutiny for such loan applications that have a lower intent score.

The Bloom filter fails for this analysis because of the false positives. The accuracy of intent scoring goes down dramatically if we just rely on the Bloom filter to find out matches in the word list. Hence, a normal ArrayList match was used to match words in the positive and/or the negative words list.

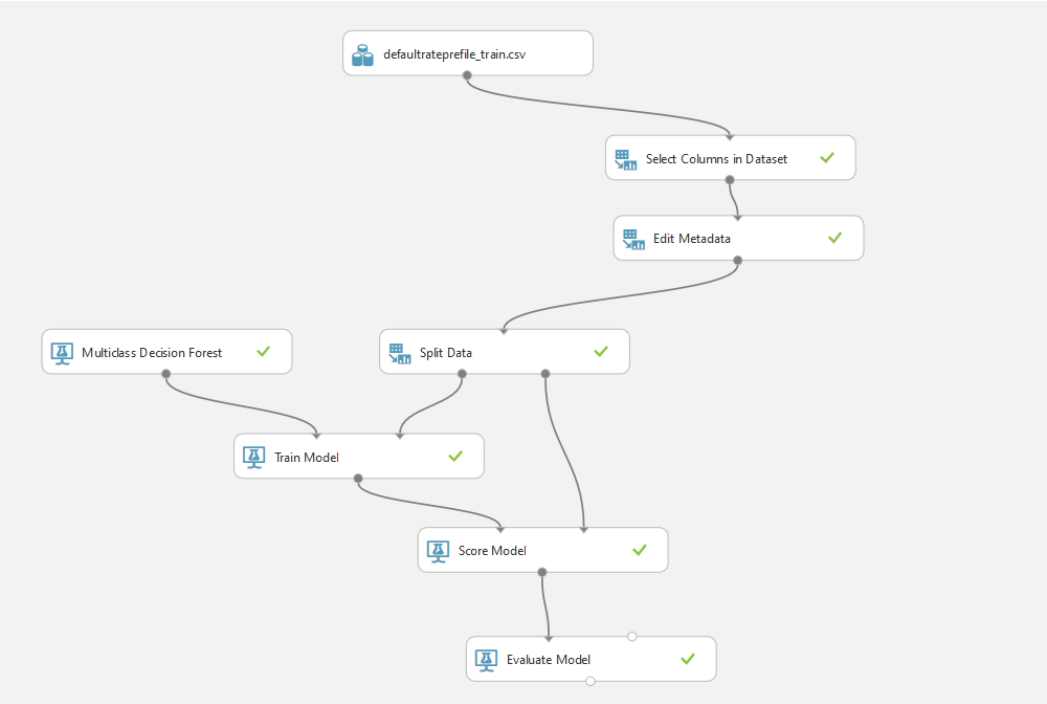
--The word list that was used was the positive and negative word list which is usually used in sentiment analysis which is why we will not be using this to predict loan defaults. Also, this is merely an average measure. Actual figures per loan application can turn out to be different. If more research is done in by linguists in this area, we can come up with another model or perform Natural Language Processing techniques in future to better calculate this intent score in future which can be used in prediction models such as the one used in the next analysis.

10. Predicting if a Loan will Default

Purpose : This analysis is a predictive analysis to determine if a particular loan will default based on features provided. The prediction algorithm used is Multiclass Decision Forest algorithm. From the previous exploratory analysis, we could determine that grade, home ownership and purpose can be important categorical features that can help in predicting if a loan will default. For increasing accuracy, we can also consider the annual income and, interest rate and dti (debt to income ratio) as numerical features that might supplement the categorical features.

We used the MR filtering pattern to filter out the relevant data i.e, loans for which a status of either a Good Loan or Default can be determined. Only the historical data was selected which could be used to train the model effectively. The current and issued loans cannot be used to train this model. The irrelevant columns were also removed as a part of this MR job.

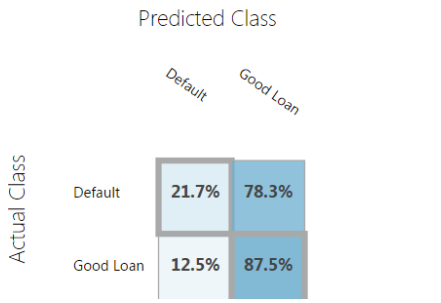
Microsoft Azure Machine Learning studio was used to develop the model.



Metrics

Overall accuracy	0.730317
Average accuracy	0.730317
Micro-averaged precision	0.730317
Macro-averaged precision	0.56332
Micro-averaged recall	0.730317
Macro-averaged recall	0.545956

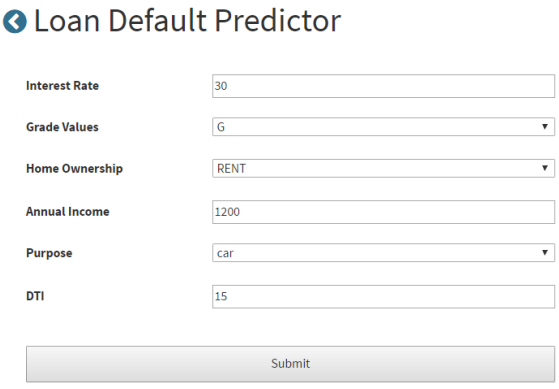
Confusion Matrix

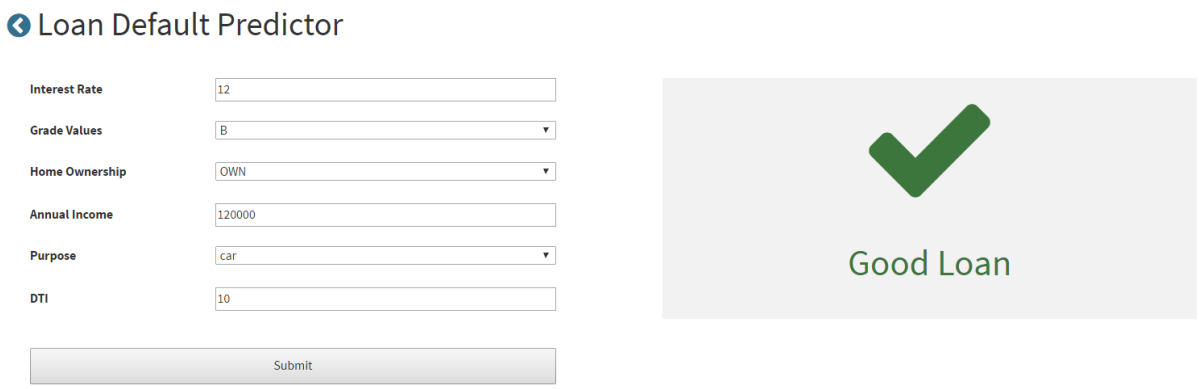


Results : The model that was computed by Azure was 73% accurate overall and could predict if a loan would default correctly 21.7% of the time.
It is understandable that this level of accuracy is very less and more research is required to come up with additional (maybe external) features to improve the prediction model.

Map Reduce patterns : Filtering patterns were used to filter out current loans for which we couldn't use as training data such as current or issued loans. This filtered data was provided as input to Microsoft Azure Machine Learning Studio to develop the model.

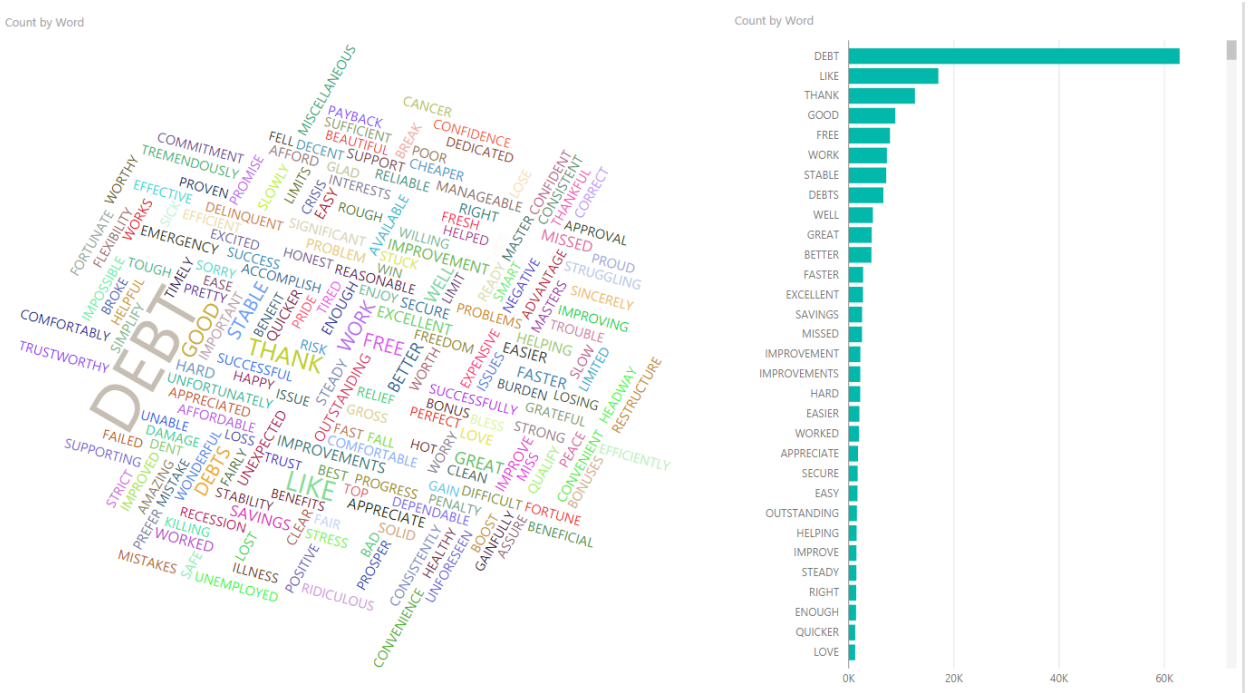
The above model was created as a web service and the output of it was used in the presentation website for display purposes.





11. Word Count Analysis for determining Hot Words

Purpose : We have categorized the intent words that are being used to calculate an intent score in the previous analysis. This analysis calculates the frequency of useful words from the loan descriptions to prepare a list of hot words. This can be used to improve the future models and also can be used by linguistic research team to increase accuracy in Natural language Techniques to improve the accuracy of the intent score that is being assigned to a loan description in future.



Future Scope

There is more work in the field of NLP that can be explored to understand the intent based on the description provided by the user. Additional externa features apart from this dataset can be explored based on demographics and financial profile of borrower to improve the accuracy of the default prediction model.