UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

SCHOOL OF INFORMATION TECHNOLOGY

STUDENT NAME: Phathutshedzo Tiny Mudau
STUDENT NO: 22783352
EMAIL: u22783352@tuks.co.za

MODULE CODE: MIT 805
MODULE NAME: Big Data Processing
ASSIGNMENT NO: 02
DATE: 30 October 2022

1 Introduction

We use big data frameworks and techniques to solve a real-world big data problem in this project. Coding the MapReduce algorithm to extract and summarize data, as well as visualizing the dataset, are examples of specific tasks. The dataset chosen is made up of historical loans from Lending Club. This is a lending club open source dataset with loans from 2007 to the fourth quarter of 2019. The data can be downloaded from the website https://www.kaggle.com/denychaen/lending-club-loans-rejects-data. According to market share, The Lending Club is one of the largest peer-to-peer (P2P) lending platforms in the United States. Since the P2P market's start in 2007, it has experienced rapid growth, sparking increased interest from consumers. There are many potentials to profit from this alternative investment choice given that billions of dollars in loans are made each year, but it is the investor's obligation to recognize the risks associated with the lending company in this market.The goal of this project is to use big data techniques to extract data from extensive records of loans provided over the platform's operational years. The dataset includes 2,650,550 issued loans, each of which comprises 159 attributes. Due to the size of the dataset (3.47GB), processing it on a system with limited processing capability is challenging. This project's major goal is to visualize numerous borrower characteristics in order to better understand how borrowers differ in this market. It also use machine learning algorithms to identify the types of borrowers most likely to default on their loan obligations.

2 Business motivation and project aim

In order to extract business insights from the loan data, the project's scope calls for the use of the MapReduce algorithm and exploratory data analysis with a big data processing framework. To quickly build a connection between the status of the loan and the characteristics of the borrowers, the MapReduce algorithm is used. Next, the features of the loans are investigated using the visualisation. Recent studies show that 3% to 4% of all loans default annually, which is the impetus behind this. This default rate is the biggest challenge for investors who want to fund the loan because the risk is assumed to be entirely on the investor and the loans issued are insecure. As a result, in order to make an informed investment decision, investors require a more granular and comprehensive assessment of the borrower's characteristics than the platform provides. Big data analytics can give investors deeper understanding of the market as a whole and the types of borrowers who use this platform to obtain loans. To give investors more interactive visualisation tools to understand this market and prevent investing in loans that are likely to default, the raw data can be summarized in the form of conversations and dashboards.

3 Technology used

The spark 2.0.7 and Hadoop 3.1.3 frameworks were used to run the MapReduce algorithm. Because of its ease of use with data analytics and machine learning models, pySpark was chosen from the spark framework. Tableau 2022.2 was chosen for data visualization because it is powerful in terms of conducting visualisation for various sorts of data, including geographical maps. Tableau can be particularly beneficial in determining borrower characteristics by geographical location because the data contains the location of the origination of the provided loans.

4 Explanatory Data Analysis

This project makes use of structured data from 2,650,550 loans with 159 features gathered between 2007 and 2019. The data types used are ordinals, categorical, and continuous. Domain knowledge was discovered to be valuable during cleaning and imputation of data because it was not possible to delete some other variables. The exploratory data analysis sought useful facts and insights regarding borrowers and loans granted through the lending club platform. The analysis will then be valuable in guiding us about borrowers' behavioral patterns as well as loan distribution. Given that it is not possible to use all 150 elements to develop predictive models, the study will assist us in deciding which variables are critical in evaluating the effectiveness of the borrowers. In order to parallelize data and process it through many local clusters, we use the Map-Reduce algorithm in PySpark. We also concentrate on data visualization with Tableau Desktop software.

4.1 MapReduce algorithm

The pyspark framework was used to implement the MapReduce algorithm to process data. We created resilient distributed dataset (RDD) for parallel processing. We were able to quickly process the dataset and explore various characteristics of the loans. By using ma, flatMap, mapValues and reduceByKey functions in spark,we were able to summarise each column(characteristics) of the loan to develop the intuition necessary for the visualisation part. Detailed results can be found on the github repository available in the link

4.2 Data visualization

In this section, we visualize data to obtain insight into the loans that have been issued. The purpose is to see various loan features to help us decide on a loan funding plan. We investigate several elements of the dataset in order to gather granular information about the loans, such as investigating the issued loans of the year and the states from where these loans originated, as well as investigating the relationship ship of loan status versus other factors.

4.2.1 Loans were made over time.

By looking at the borrower's credit score, credit history, debt-to-income ratio, and the amount requested, the platform assesses the borrower's creditworthiness. All of these traits are assessed using the credit grade metric, which uses an alphabet spanning from A to G with the letter A standing for low risk and the letter G for high risk. Five sub-grades, ranging in risk from low to high, are included in each of these grades. The sub-grades are given to further separate various risk profiles and provide more incremental information on the risks related to the borrower.

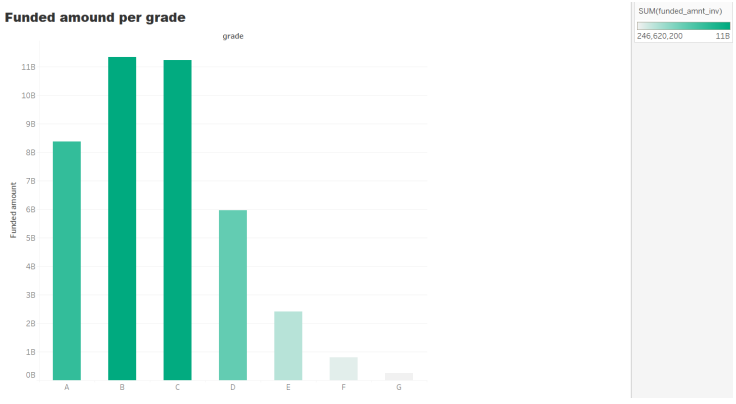

Figure 1: Loans made over time.

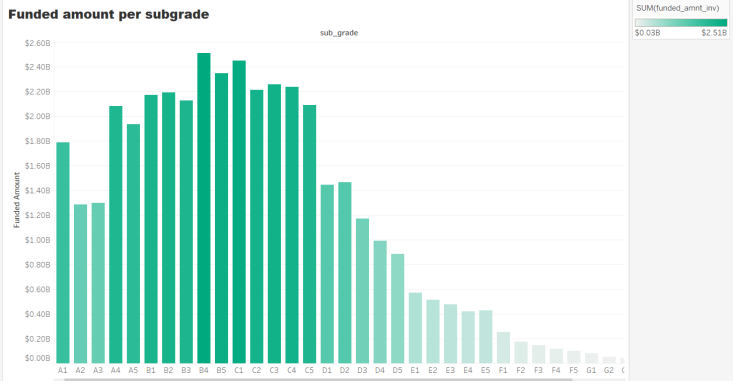

Figure 2: Loans made over time.

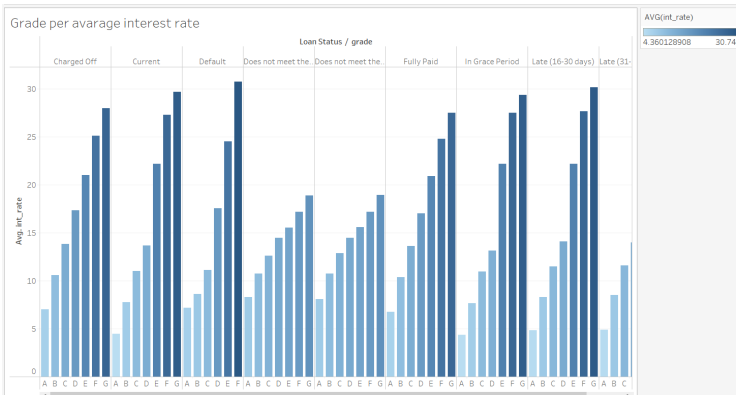

Figure 3: Loans made over time.

4.2.2 Loans issued by term, verification status, and loan status.

Figure () shows that loans with current, completely paid, and charged off statuses dominate the data. As a result, the other loan statuses do not disclose any more information, and it is worthwhile to disregard them when conducting further analysis.

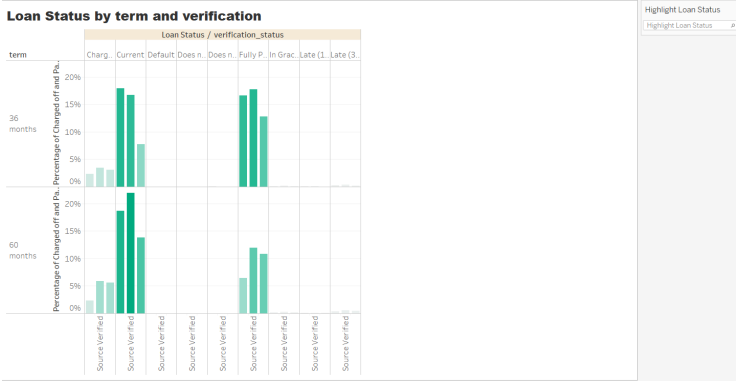

Figure 4: Loans made over time.

Figure 5: Loans made over time.

4.2.3 Loans made by state.

From Figure 2, we can see that majority of the clients who are applying for loans comes from California (CA), Texas (TX) Florida (FL), New York (NY) and Illinois (IL). Interested investor can therefore focus on drilling down borrowers characteristics from these states in order to have granular information. An interesting investigation will be do

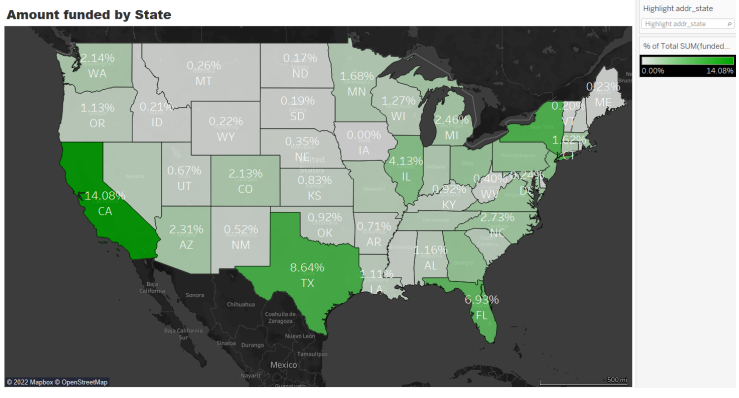

Figure 6: Loans made over time.

4.2.4 Loans made according to purpose.

The purpose of the loan must be understood, just like with traditional financial intermediaries, in order for expert judgment to determine if it is possible to grant the loan for the specified purpose. Figure below displays a word cloud of the most popular reason for which loans are requested. It is clear that most people use loans to pay off their obligations.

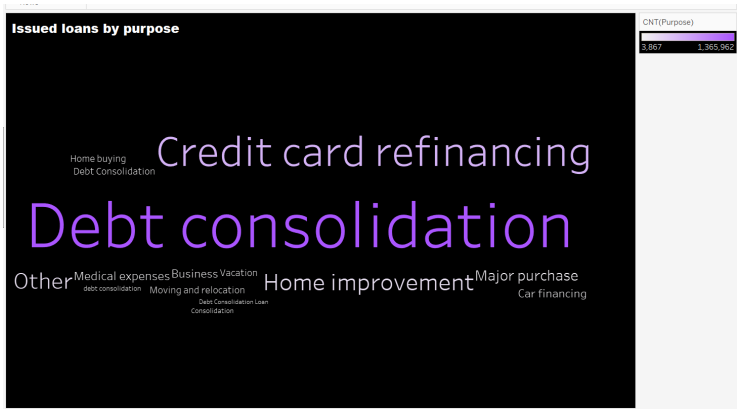

Figure 7: Loans made over time.

5 Conclusion

In order to learn more about the characteristics of borrowers and loans that are likely to be repaid on time and those that might default, we use big data analytics to the lending club data. The spark architecture has offered a useful tool to quickly process the enormous loan record. Using the spark MapReduce architecture, we can quickly summarize the data and begin to construct an intuitive understanding of the loans available on the Lending Club platform. A solid foundation for data visualization is provided by the MapReduce findings. Our visualization demonstrates how three statesCalifornia, Texas, and Floridacontribute more to the overall loan volume granted by Lending Club, with the majority of the loans going toward debt consolidation. The data visualization also demonstrates that bigger loan amounts have higher default risk, loans with shorter terms 36 months show higher default rates than those with longer terms 60 months and some loans with particular sub-grades have higher default rates than loans with other sub-grades. With this knowledge, potential investors should concentrate on loans with 60 month periods and grades B and C, respectively.