**ITCS 6100**

**Big Data Analytics for Competitive Advantages**

**Group 15**

**canvas-sample-datasets**

**Team Members**

1. Aakshi Soni (801275487)

2. Akshay Narkhede (801275760)

3. Saloni Avhad (801322583)

4. Santosh Reddy (801306775)

5. Soumik Paul (801308500)

**Dataset**

For understanding and exploration of the loan repayment status we have chosen this dataset. The dataset consists of 21 column and around 40000 rows.

S3 PATH s3://projectbucket15/canvas-loan-dataset.csv

**AWS Services Used**

1. AWS S3
2. AWS Glue
3. AWS Athena
4. AWS QuickSight

**About Data**

Research objective is to predict, whether a customer will repay a loan or not. The dataset contains complete loan data for all loans issued from 2007-20011 including the current loan status and latest payment information. Target column for this dataset is loan\_status. The dataset consists of ~40000 rows, 21 features columns.

**Creation of Bucket and storing the dataset into the bucket**

Bucket “group15bucket” has been created and the dataset in csv format was added.

Graphical user interface, text

Description automatically generated

Graphical user interface, text, application, email, website

Description automatically generated

**Data Preparation using AWS Glue**

**Step 1:** AWS Glue is used to fetch the data from S3 to AWS Glue.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, email

Description automatically generated

Properties of the table

Graphical user interface, application, Teams

Description automatically generated

Table Schema

Graphical user interface, application, Teams

Description automatically generated

**Step 2:** Previewed the data on AWS Athena

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, application

Description automatically generated

**Step 3**: Run few SQL queries on AWS Athena for understanding the datasets.

1. Avg of loan amount

Graphical user interface, text, application, email

Description automatically generated

1. Avg of interest rate

Graphical user interface, text, application

Description automatically generated

1. Avg of Loan Term

Graphical user interface, text, application, email

Description automatically generated

**Data Visualization on Quick Sight**

The dataset has been visualize on AWS Quick Sight to get a better understanding of the data.

Count of output column (loan\_status)

Chart, bar chart

Description automatically generated

Average of interest rate vs count of sub grade

Table, treemap chart

Description automatically generated

Count of sub grade, average of interest rate and average of interest rate by sub grade.

Chart, scatter chart

Description automatically generated

Installment vs count of issued on by date column

Chart, scatter chart

Description automatically generated

Count of records by home ownership

Graphical user interface, chart, application, pie chart

Description automatically generated

interest rate by date

Graphical user interface, application

Description automatically generated

**Quick Sight Dashboard**

<https://github.com/SoumikPaul108/BigData_Final_Project_Group15/blob/main/Sheet_8_2022-11-14T21_28_53%20(1).pdf>

Chart, scatter chart

Description automatically generated

## ****Analytics, Machine Learning****

## **XGboost**

## Graphical user interface, text, application, email Description automatically generated

## Random Forest

## Table Description automatically generated with low confidence

## ****Evaluation and Optimization****

## ****A picture containing graphical user interface Description automatically generated****

## ****Graphical user interface, text, application Description automatically generated****

## ****Table Description automatically generated****

## ****Table Description automatically generated****

## ****Results****

## **XGBoost**

## **Table Description automatically generated**

## ****Graphical user interface, text, application, email Description automatically generated****

## **Random Forest**

## **Table Description automatically generated**

## ****Graphical user interface, text, application, email Description automatically generated****

## ****Future Work****

**What was unique about the data?  Did you have to deal with imbalance?**

The entire dataset has around 40k rows and 21 columns. The dataset was split into two halves. As a result, we had to conduct an inner join operation on the datasets in order to operate on only one.

**What data cleaning did you do? Outlier treatment?  Imputation?**

First, we determined the percentage of null values in each column, ordered in descending order of missing values; "employment length" was the sole column with 2.7% of the null value. Second, we view the null values and drop them entirely.

**Did you create any new additional features / variables?**

We have not added any additional features. Only the features and functionality described in the instructions and AWS labs have been preserved in the project.

**What was the process you used for evaluation?  What was the best result?**

Our machine learning model was constructed using the XGboost and Random Forest algorithms. In the case of XGboost, we reached 86% accuracy, and in the case of Random Forest, we also achieved 86% accuracy. Using both techniques, we were able to obtain the same accuracy.

**Is there Bias in your work? What were the problems you faced? How did you solve them?**

The only challenge we had was that all the text columns' data became corrupt when we uploaded our information to AWS Athena via AWS Glue, while all of the integer and double columns' data remained in pristine condition. Initial inquiries were successful; however, the dataset began to display erroneous data after a few AWS Athena queries. The dataset, however, worked flawlessly when visualized using AWS Glue.

**What future work would you like to do?**

To determine how many of the current loans were paid off, defaulted on, or even charged off, we may utilize an updated data frame that includes the numbers for the next three years (2007-2011). These new data points can then be utilized for forecasting or for developing new models.

In order to improve the model's ability to anticipate competent borrowers, we may wish to look into this matter more since the algorithm places about 36% of non-defaulters in the default class.

**Instructions for individuals that may want to use your work**

LendingClub must use caution when finding potential borrowers who meet specific requirements. Borrowers who do not own a home, for example, and apply for a small company or wedding loan may have a bad combination that leads to the borrower defaulting on a future loan.

LendingClub must be aware that low-graded loans are undoubtedly more likely to default. They must be prepared to collaborate with these borrowers to secure appropriate and timely payments. Reduced interest rates or installment payments for these consumers might be beneficial.