**CST\_2105\_300**

**DATA SCIENCE FOUND. FOR BISI**

**TEAM 11:**

Hetkumar Patel

Dhruv Gadhiya

Jay Patel

**Project Title:** Loan Repayment Prediction

**Business Understanding:**

* An essential responsibility in the banking and financial industry is the prediction of loan payback. Building a predictive model that can correctly forecast whether a borrower will return their loan amount which they have taken for their business's aim. The following are some ways in which this can benefit banks and other financial institutions:
* Financial organizations can prepare and have better manage risk management which helps them to reduce their losses by anticipating the possibility of loan defaults.
* Models that forecast how much a loan will be repaid can assist banks in making better choices regarding who to lend to and under what conditions.
* Customer service: By identifying borrowers who are at risk of default and collaborating with them to find a solution, predictive models can also assist banks in offering better customer service.

**Background:**

* So, we got this dataset from Kaggle. Basically, this dataset is related to loan details taken by the borrower.

**Problem Statement:**

* So here we try to develop a model and classification model to accurately predict whether the borrower fully paid the loan amount or not.

**Observations:**

* Loan repayment prediction is a machine learning problem that involves predicting whether a borrower will repay their loan in full or not. In this dataset, there are 18 features (including the target variable) and 9578 observations. The features include the purpose of the loan, the interest rate, the borrower's credit score, and other financial information.

**Data Capture and Explanation:**

* Based on the given data, there are several variables that describe different aspects of loan applicants, such as their credit policy status, purpose of the loan, interest rate, installment amount, log of annual income, debt-to-income ratio (DTI), FICO credit score, days with credit line, revolving balance, revolving utilization rate, inquiries in the last 6 months, delinquencies in the past 2 years, public records, and whether or not the loan was fully paid.
* The data seems to be a sample of loans, and it appears that some of the loans were not fully paid back. The data could potentially be used to build a model that predicts the likelihood of a loan being fully paid back based on the available features.

**Data Exploration (clustering, visualization, etc.):**

* We found this dataset from Kaggle, originally it was the data of bank which have the details of all the borrower who took the loan for specific purpose from bank.
* It has total 8 features including credit policy, purpose, interest rate, installment, debt-to-income ratio, FICO Credit score, number of days the borrow has the credit line, revolving balance, revolving line utilization rate, number of inquiries, any public derogatory public records and loan is fully paid or not.

**Data Preparation:**

* In machine learning data preparation is crucial part because insufficient data preparation will result the model in underfitting.
* It is a process of transforming raw data into a clean, organized, and useful format for ML models.
* The basics steps of data preparation:

1. First, we do the data clean to remove the unnecessary data from the dataset which includes removing duplicates, correcting errors, and many more.
2. Second, we do the handling of missing value, but in our dataset, there is no null values.

**Feature Engineering:**

* The process of choosing, manipulating, and extracting pertinent features from unprocessed data to produce a dataset that is appropriate for machine learning algorithms is known as feature engineering.
* It is a crucial phase in the machine learning workflow since the caliber and applicability of the features used to train a model have a significant impact on how well it performs.
* The purpose of feature engineering is to develop features that extract irrelevant or superfluous material from the data while capturing crucial information. Many methods, including scaling and normalization, imputation of missing values, feature selection, feature transformation, and feature extraction, may be used in this procedure.
* In conclusion, feature engineering is a critical step in the machine learning pipeline and necessitates careful thought and experimentation to create the best set of features for a given problem.

**Modeling and Pipelines Creation:**

* Here we the use the 2 models algorithm (SVM and KNN) and 1 classification (Decision Tree Classification):
* Support Vector Machines:

Machine learning algorithms such as Support Vector Machines (SVMs) offer powerful capabilities for classification and regression. By maximizing the margin between the hyperplane and the closest data points, it finds the best hyperplane that separates data points into different classes.

* KNN:

It is a machine learning algorithm used for classification and regression tasks. It does not make any assumptions about the distribution of the data and does not require any training or learning phase.

* Decision Tree:

It is a type of supervised machine learning used to categorize or make predictions based on how a previous set of questions were answered.

**Hyperparameter Tuning:**

* To do the hyperparameter tuning here we use GridSearchCV to find the estimators for Decision Tree model.

**Results Interpretation:**

* Machine learning classification is the task performed in the code. Based on various factors, such as Credit Policy, purpose, installment, interest rate and dti, the task is to predict whether the borrow fully paid the loan amount or not.
* Here we use the ROC Curve to analyze the results. The ROC curves for each model also show the trade-off between the false positive and true positive rates based on different classification thresholds. AUC is a measure of the model's performance based on the area under the ROC curve (AUC).

**Screenshots:**

**Chart, histogram

Description automatically generated**

**Chart

Description automatically generated**

**Chart, histogram

Description automatically generatedChart, bar chart

Description automatically generated**

**Check duplicates**

**A picture containing graphical user interface

Description automatically generated**

**Normalization**

**Graphical user interface, text, application, email

Description automatically generated**

**Models and classifications**

**Graphical user interface, text, application

Description automatically generated**

**Hyperparameter tuning**

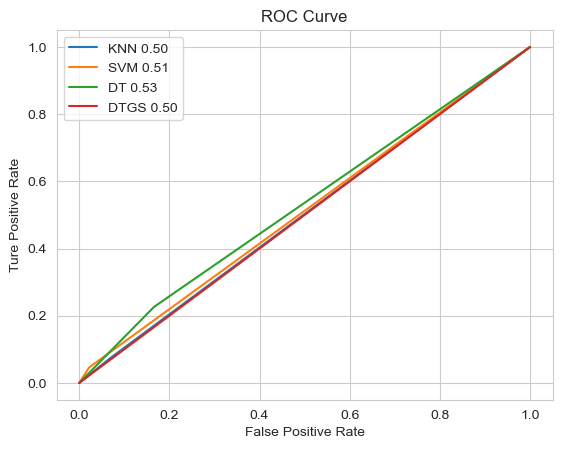
**Graphical user interface, text, application, email

Description automatically generated**

**ROC Curve result**

**Graphical user interface, text, application, email

Description automatically generated**

****