A picture containing text

Description automatically generated

Assignment-1

BIG DATA

INSTRUCTOR**:** GRAHAM WALL

CLASS**:** CBD 2214

TEAM MEMBERS: LEKSHMI CHANDRAN SHEELA

PARVATHY VYSAKH

PATRICIA ADOLPH

POOJA SELBY

SUCHITRA CHANDRASHEKHARAN

**Content**

1. Introduction
2. Data Description
3. Exploratory Analysis
4. Data Pre-Processing
5. Data Segmentation
6. Model Building
7. Comparison and Conclusion

**Introduction**

Banks are financial intermediaries that play a crucial role between savers and borrowers. Bank manages deposits flowing in and loans flowing out to a bank for the soothing functioning of society and the market. Different criteria like the credit score, age, monthly income, etc. are analyzed to decide whether a loan should be granted or not. Credit risk for a loan is computed by the likelihood that the borrower will repay the principal amount and the interest for a loan. Banks perform an in-depth credit background check before approving any credit. Many factors affect credit risk modeling, and they are the financial status of the borrower, the consequences of a default, credit extension size, past default rates trends.

Three factors that have a stronger association with credit risk modelling are:

1. Probability of default
2. Loss given default
3. Exposure at default

Credit losses (a debt that a firm is unlikely to be able to pay back) on financial instruments (e.g., loans, bonds, etc.) were considered by the "loss model" during the 2008 fiscal crisis, and as a result, assets were classified as financially impaired when credit quality degradation and considerable time delays were issues. Even if the organization expects to be paid in full but later than when legally due, a credit loss occurs since projected credit losses include the quantity and timing of payments. Expected Credit Loss is the probability-weighted estimate of credit losses throughout the projected life of a Financial Instrument. It can be calculated using one of two methods:

1. Allowance matrix - based on internal, earlier credit loss data and delinquent payment receivables of an entity.
2. Valuation model - based on default probabilities.

Larger institutions use the PD (probability of Default) analysis approach to determine their expected loss. The most common way to analyze a PD is to look at past-due debts.

Here, a model is built to predict whether the borrower will pass 90 days due to delinquency by learning the probability of default, using machine learning algorithms.

The following are the steps involved in PD prediction.

1. **Data Description**

The dataset for credit risk modelling has been downloaded from the Kaggle dataset “Give Me Some Credit”. The dataset contains 11 features with 150,000 sample data which is to be analyzed to mark risk assessment scores for loan applicants.

1. SeriousDlqin2yrs: 90 days past due delinquency experienced by a person.
2. Revolving UtilizationOfUnsecuredLoans: Total balance on credit cards and personal lines of credit except for real estate and no installment debt like car loans divided by the sum of credit limits.
3. Age: Age of borrowers
4. NumberOfTime30-59DaysPastDueNotWorse: Number of times borrower has been 30-59 days past due but no worse in the last 2 years.
5. DebtRatio: Monthly debt payments, alimony, living costs divided by monthly gross income.
6. MonthlyIncome: Income of the borrower.
7. NumberOfOpenCreditLinesAndLoans: Number of Open loans (instalment like car loan or mortgage) and Lines of credit (e.g., credit cards)
8. NumberOfTimes90DaysLate: Number of times borrower has been 90 days or more past due.
9. NumberRealEstateLoansOrLines: Number of mortgage and real estate loans including home equity lines of credit.
10. NumberOfTime60-89DaysPastDueNotWorse: Number of times borrower has been 60-89 days past due but no worse in the last 2 years.
11. NumberOfDependents; Number of dependents in family excluding themselves (spouse, children, etc.).
12. **Exploratory Analysis**

It is a clever idea to first comprehend the data and then strive to extract as many insights as possible. With the aid of summary statistics and graphical representations, exploratory data analysis can be performed. It is an important process of doing early investigations on data to find patterns, spot anomalies, test hypotheses, and check assumptions.

These are the inferences we got from the data set:

* Dimension of the data – 150000 rows and 12 columns.
* There are no duplicate rows in the dataset.
* Presence of unbalanced data. Unbalanced data will cause the supervised learning algorithm to pay too much attention to most classes and reduce the classification performance.
* The minimum age is 0, which should be removed.
* Data distribution plot of DebtRatio, NumberOfTime30-59DaysPastDueNotWorse, NumberOfTime60-89DaysPastDueNotWorse, NumberOfTimes90DaysLate, NumberRealEstateLoansOrLines, RevolvingUtilizationOfUnsecuredLines are abnormal, i.e., there are some extreme values that affect the presentation of the distributed image.
* From the distribution curve for customer income and age both variables are normally distributed, which is in line with the assumptions of statistical analysis.

1. **Data pre-processing**

Data from the real world is always unclean with missing values, errors and can always have noise, outliers, etc.Data pre-processing is the technique to transform real-world data into an understandable format. Our dataset has been pre-processed using two methods:

1. Handling missing values

Missing values should be managed before actual modelling because they lead to faulty decisions. Before starting our credit risk modelling our first step is to process the missing values. Following are some methods used for handling missing values:

* Delete rows with missing values.
* Manual filling based on similarities between samples.
* Filling missing values using a correlation between variables.
* Deleting redundant data.

Here missing values were present in two features, MonthlyIncome, and NumberOfDependents, For Monthly income they are replaced using random forest regressor. For Numberofdependents median imputation is used to manage missing values.

1. Outlier Detection and Processing

An outlier is a data that stands significantly different from other observations in the dataset. The PD dataset contains outliers in all features, and it is removed using a boxplot. If an observation in the dataset for each features is above a particular threshold point has to be removed. The following are the thresholds that we observed from the boxplot and they are removed from the dataset.

* Age < 0
* Debt Ratio > 8000
* MonthlyIncome > 50,000
* Numberof Dependents > 10
* NumberOfTime30-59DaysPastDueNotWorse > 20
* NumberRealEstateLoansOrLines > 30
* RevolvingUtilizationOfUnsecuredLines > 3

After removing all the outliers and extreme values we have 148771 rows and 12 columns.

1. **Data Segmentation**

Data segmentation is the method of dividing and grouping similar data into distinct groups. Verify our dataset’s fitting effect on the model, the whole dataset is segmented into a training set and test set. Then used the training set to train the model and testing set to test the efficiency of the model.

After removing the outliers, we have 148771 rows, 11 columns in the data set. The data is split in the ratio of 80:20 for training and testing, respectively. After splitting independent train set is 119016 rows, 11 columns and the test set is 29755 rows and 11 columns. The dependent train set is 119016 rows, and the test set is 29755.

1. **Model Building**

During the initial stages, we were unaware of which model would give us more accuracy, we used three models to train and test the data.

* **XGBoost**

Extreme gradient boosting or XGBoost is an ensemble Machine learning method that comes under the Gradient Boosting framework. It is an advanced implementation of the Gradient Boosting algorithm designed to boost the speed, performance, and accuracy of the model. This technique has the advantage of preventing overfitting by using a variety of regularizations to even the final weights. XGboost is an ensemble model which takes too much time to train the data. So to save time, instead of training the model again and again, during the next model fitting, we saved the best estimator after training the data to a pickle file named xgb\_model. pickle file, so that the trained model can be reused. Then predict the probability of the target variable using independent variables in both the train and test data set.

ROC or **Receiver Operating Characteristics Curve is a performance assessment measure for any classification model at various thresholds. The area under the curve (AUC) shows how well the model distinguishes between positive and negative classes. The greater the AUC, the better. The ROC curve we got from this model is shown below.**

Chart

Description automatically generated

Figure 1: ROC curve for XGBoost

The image clearly shows that our model's discriminating ability to discriminate between positive and negative classes is good.

Confusion Matrix is another machine learning classification performance metric. Using the real and predicted target variable confusion matrix is created. From the confusion metrics of the trained XGBoost model, we got 110921 values as True Positive and 1312 values as True Negative. Similarly, for the tested XGBoost model, 27647 as True Positive and 227 as True Negative. Using these real and predicted values accuracy, precision, and recall of the model in test and train data is calculated as shown below. Table

Description automatically generated

Figure 2: Performance Evaluation using classification Report

The overall accuracy of this model is 94%.

* **Random Forest**

  Random Forest is categorized as a supervised learning algorithm. It is a collection of decision trees and is a powerful technique to use for model prediction and behavior analysis. Each tree in the forest represents an instance of classification of data as input into the random forest. These instances are combined to get more accurate and precise prediction results. Since our credit risk modeling dataset is vast random forest model runs efficiently with a lower risk of overfitting.

The below picture shows that there is a great imbalance in the target variable as class ‘1’ has more observations than class ‘0’.

Chart, bar chart

Description automatically generated

Figure 3: Classes in target variable

Imblearn approaches are ways for generating a data collection with an equal proportion of classes. This sort of data collection would allow the prediction model to generalize effectively. Up-sampling and Down-sampling are the two major methods for dealing with an unbalanced data collection. Imbalanced classification involves developing predictive models on classification datasets that have a severe class imbalance.

Synthetic minority oversampling (SMOTE) is one approach to address imbalanced datasets by oversampling the minority class. This is done by duplicating instances in the minority class but does not supply any new information to the model. Instead, new instances may be created by synthesizing old ones.

Since this model takes too much time to train, instead of training the model each time, we save the fitted model as a pickle in the ‘rfc\_model.pickle’ file, so that the trained model can be reused. Like XGBoost, we predicted the probability of the target variable using independent variables in both the train and test data set using a random forest classifier.

Between the true positive and false positive rates, a ROC curve is created, and the area under the curve (AUC) is determined to be 0.86, which is excellent. From the confusion matrix, True Positive is 92527 and True Negative is 5479 for the train data set, and for the test data set it is 22870 and 1321 respectively. Using these real and predicted values accuracy, precision and, recall of the model in test and train data is calculated.

Table

Description automatically generated

Figure 4: Performance evaluation of Ramdom Forest

The overall accuracy of this model is 81%, lower than XGBoost.

* **Logistic Regression**

Logistic Regression is a technique used to calculate the relationship between two variables (dependent variable should be categorical). It measures the relationship between the dependent variable with one or more independent variables using a logistic function. For a binary logistic model, the dependent variable with two possible values labelled as’0’ or ‘1’. Logistic Regression is easier to implement and very efficient to train We have predicted the probability of target variable using independent variables in both the train and test data set using Logistic Regression. A ROC curve is created between the true positive and false positive values. The area under the curve (AUC) is resolute as 0.86. The confusion matrix for the trained logistic regression model shows that True positive and True negative values are 90045 and 5806 respectively.

Table

Description automatically generated

Figure 5: Performance evaluation of Logistic regression model

The overall accuracy of the logistic regression model is 80%.

**Comparison and Conclusion**

Our goal was to build a model which predicts whether the borrower will pass 90 days due to delinquency by learning the probability of default. By comparing the accuracy from all the models, XGBoost gives more accuracy than others which is 93 percent. Details of each model are shown in below diagram. XGBoost can be used to predict multiple problems instead of using several algorithms. It enhances the performance and accuracy of the classification and regression problems. It ensures quick execution of multiple cycles while tuning hyperparameters. From the confusion matrix, the True positive and True Negative for trained XGBoost model is 110921 and 1312 respectively which is comparatively higher than Random Forest and Logistic Regression.

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Figure 6: Comparison of three models

The following are the conclusion and finding from our analysis and modeling.

* + Three models (XGBoost, Random Forest, and Logistic Regression) were built to predict the Probability of Default.
  + By comparing the accuracy from all the three models, it is evident that XGBoost gives more accuracy than others
  + Selected Model - XGBoost: Highly Flexible, High Accuracy rate, uses the power of parallel processing and Faster.
  + Banks can use this PD model in order to predict Excepted Credit Loss.

**References**

Bhalla, D. *A complete idea to Credit Risk Modelling. Listen Data*

<https://www.listendata.com/2019/08/credit-risk-modelling.html>

Chen, T., & Guestrin, C. (n.d.). *XGBoost: A Scalable Tree Boosting System*.

<http://dmlc.cs.washington.edu/data/pdf/XGBoostArxiv.pdf.>

Koehrsen, W. (2018, August 31). *An Implementation and Explanation of the Random Forest in Python*. Medium.

[https://towardsdatascience.com/an-implementation-and-explanation-of-the- random-forest-in-python-77bf308a9b76](https://towardsdatascience.com/an-implementation-and-explanation-of-the-%20%20%20%20%20random-forest-in-python-77bf308a9b76)

*What is Logistic Regression?* Statistics Solutions. (2021, May 4). <https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-logistic-regression/>

*Handling Missing Values and Outliers - 1457 Words: Assessment Example*. Free Essays. (n.d.).

<https://ivypanda.com/essays/handling-missing-values-and-outliers/>

Sharma, L. (2019, November 3). *What is Pickle in python ?* Medium. <https://medium.com/@lokeshsharma596/what-is-pickle-in-python-3d9f261498b4.>

*What is XGBoost Algorithm - Applied Machine Learning*. DataFlair. (2021, March 8).

<https://data-flair.training/blogs/xgboost-algorithm/.>

**Appendix**

The following are the python code and results extracted from the analysis.

