Technical Report for Machine learning and Pattern Recognition

Subject Code - B9DA107

Group Members and ID:

Morenikeji Ifeoluwa Egbeyeye (10568950)

Okhunlun Onomeasike Andrew (10563664)

Parimal Khushal Bangre (10579654)

Introduction

Subscribing to bank loans is one of the ways people in various countries pay for purchases. Financial institutions, banks, government organizations and even loan companies. Some of these loans assist in buying houses like mortgages, in getting cars, paying for school fees, and a list of other advantages. But loan giving comes with its own risk. Financial institutions must consider the risk involves before giving out loans and find ways to minimize this risk. As a result, various ways and techniques are being developed to determine the outcome of a loan before it is given out and one of such ways is machine learning.

Whether adequately labeled data is used (supervised learning) or not, machine learning is the science that allows computers to detect patterns within data and create classification and prediction models (unsupervised learning). As earlier stated, machine learning is divided into two: Supervised learning and unsupervised learning. Supervised learning is the form of machine learning that requires you to "specify the output" to use it. It takes advantage of well-labeled data. After then, a new collection of data known as the test set is introduced to the model, which results in a result. Logistic regression is a sort of supervised learning that focuses on making predictions about situations that require a probability estimate as an output. We intend to use logistic regression to predict the prepayment risk of vehicle loans because it is used to determine the probability of any event occurring. For our project, we’ll utilize logistic regression, a supervised machine-learning approach for estimating the likelihood of an event happening. This will help banks minimize the loss in profit caused by prepayment risk. Prepayment risk is defined as "the chance that a borrower may pay off the loan before it matures," It can result in a loss of interest for banks and often requires them to reinvest the prepaid funds at a lower rate of return. In this research, we will focus on vehicle loan prepayment and develop a model that can predict whether a subscriber would return the loan before its maturity date. The following questions will be attempted to be answered by this initiative. Can we accurately anticipate the likelihood of clients or borrowers failing on their first EMI of a vehicle loan on the due date?

This classification problem's target variable is "loan default." It indicates whether the customer is in default.

Data

For our analysis, we will be using a dataset called CarLoanDefaulters. Our dataset has (31818 rows, 41columns)

Exploratory data analysis (EDA)

Exploratory data analysis is an approach in statistics used to analyze datasets to summarize its main characteristics. Usually, it is a model used to get the general idea of the data, interpreting what the data can tell us beyond modeling or hypothesis testing. Sometimes EDA involves virtualization, finding outliers in the data, finding missing values etc. in python commands such as shape, info, describe, histogram etc. Before analyzing or modeling a dataset it is important to carry on EDA on the dataset. On performing EDA on our dataset, we found 7661 missing values in the Type of employment column. Because they are different ways on dealing with missing data, we first calculated the percentage of the missing value and it showed 3.9% which is a negligible amount, we then dropped all the missing values. We also did a histogram visualization on our dataset to show the distribution of our dataset. More on the EDA and result are in the python code files.

Choice of dependent and independent variables and selection of algorithm

After conduction EDA on our dataset, we were able to have a better idea of our data. Our problem definition involves creating a model than can be used to predict if a customer can repay the loan given. Hence loan_default is best as our dependent variable. Our loan_default variable has classes 0 for default and 1 for repaid. Based on this, Logistic regression is the best algorithm for our model. Logistic regression an algorithm used in machine learning (Borrowed from statistics) is the go-to method for binary classification problems. (Classes with two values 0 and 1)

Data Preparation

After selecting our dependent variable or output and our model algorithm, the next step for us is our data preparation. For data preparation we first printed out the list of numerical variables, and then for categorical variables. We found out that some of our categorical variables like Credit_History and Avergae_loan_period had string so using the command

```
dataset['Average_loan_period'] = dataset['Average_loan_period'].str.replace('yrs ','.',regex=False)

dataset['Average_loan_period'] = dataset['Average_loan_period'].str.replace('mon','',regex=False).astype(float)
dataset['Credit_history'] = dataset['Credit_history'].str.replace('yrs ','.',regex=False)
dataset['Credit_history'] = dataset['Credit_history'].str.replace('mon','',regex=False).astype(float)
```

We converted them to string and used the dataset[categorical_feature_columns].head()
To view the result

Also, for type of employment we encoded them to salaried as 0 and self-employed as 1 for easier analysis. As part of our data preparations, we plotted distribution of classes to target variables and we saw that there was imbalance in the classes and as a solution to it, we over sampling for the classes one (1). we used the command

```
dataset = dataset.sample(frac=1)
loan_default_1 = dataset.loc[dataset['loan_default'] == 1]
loan_default_0 = dataset.loc[dataset['loan_default'] == 0]

normal_distributed_df = pd.concat([loan_default_1, loan_default_1, loan_default_1, loan_default_0])
```

```
# Shuffle dataframe rows
new_df = normal_distributed_df.sample(frac=1, random_state=42)
new_df.head()
```

Feature Selection

This strategy is designed to minimize the number of input variables to those that are thought to be the most valuable to a model in predicting the target variable. Some predictive modeling issues contain a huge number of variables, which can slow down model construction and training and necessitate a lot of system memory. Additionally, certain models' performance can suffer when input variables that are unrelated to the target variable are included. Part of our feature selection included dropping unnecessary variables like data of birth, Bureau_Score_Description and Provider_ID. After dropping these features, it reduced our data set from 41 features to 37 columns and 225493 rows.

Model Development and Evaluation

We have initially mentioned that we would be using logistic regression for our modeling. For our model evaluation, we first divided our dataset into train and test. Using same data for both training and test for model would lead to over fitting in machine learning. In machine learning (Borrowed from statistics), overfitting is the production of an analysis that corresponds too closely or exactly to a particular set of data and may therefore fail to fit additional data or predict future observations reliably. We used train-test-split of 70 to 30 ratio. 70 for training and 30 for testing our model. And then we defined the values to evaluate our model

Confusion matrix
Accuracy score
Precision
Recall
F1 score
ROC AUC score.

The model provided the following result:

```
Confusion Matrix:
 [[38291 14681]
 [24434 19623]]
Accuracy:  0.5968730997949067
Precision:  0.5720324160447762
Recall:  0.44540027691399775
f1 score:  0.5008358749888339
```

```
roc_auc_score:  0.584126929969496
```

Model Comparison

Stochastic Gradient Descent (SGD) is a simple yet very efficient approach to fitting linear classifiers and regressors under convex loss functions such as (linear) Support Vector Machines and Logistic Regression. Strictly speaking, SGD is merely an optimization technique and does not correspond to a specific family of machine learning models. It is only a *way* to train a model.

After applying SDG to optimize our model we were produced the following result

```
Confusion Matrix:
 [[38394 14578]
 [24367 19690]]
Accuracy:  0.5986251533046821
Precision:  0.5745885374109957
Recall:  0.4469210341148966
f1 score:  0.5027768911586338
roc_auc_score: 0.5858595203044467
CodeText
```

Comparing both model result (With optimization and without optimization), There isn’t much difference. Hence for future analysis, we would apply other algorithm such as decision trees, random first and them compare the results.

References

Kaggle.com. 2021. *L&T Vehicle Loan Default Prediction*. [online] Available at: <https://www.kaggle.com/mamtadhaker/lt-vehicle-loan-default-prediction> [Accessed 4 November 2021].