Predicting Credit Card Approval of Customers using Machine Learning Algorithms

CIND 820 – Big Data Analytics Project

Supervisor: Dr. Ashok Bhowmick
Date of Submission: February 14, 2022

Prabhkiran Kang
Student Number: 501068149

Ryerson University

Prabhkiran Kang
501068149

Table of Contents

Abstract…….Page 3

Introduction……..…….Page 4

Literature Review……..Page 5

Data Description……...…….Page 7

Data Source……...…….Page 7

Data Description……..……..Page 7

Table 1: Application_record.csv data description**……..Page 8**

Table 2: Credit_record.csv data description**…….……..Page 9**

Approach Overview…….…….Page 9

References……Page 10

Abstract

Banking industry contains large volume of data on customer's information and receives a large number of applications on the daily basis for a credit card request. As the number of applications to obtain a credit card increases, it becomes hectic to approve a credit card request through a manual process as it takes more time and effort. It is also easy to make errors with a manual process. Thus, it is important to automate a process to increase the efficiency and decrease the response time. By automating a process, banks would be able to classify and divide the applicants into categories of 'Good Client' and 'Bad Client'. This way bank can decide whom to approve for a credit card and whom to reject. This also lowers down the risk of any future credit defaulters, saving financial institutions lot of money. Process of making a decision can be automated using machine learning algorithms.

In this project my goal is to use the historical data and machine learning algorithms to predict if a customer's credit card application would be accepted or rejected. I would be applying 5 different classification algorithms to the dataset to find out which model gives the most accurate prediction results. The theme of my choice is Classification. I would be using a dataset from Kaggle Inc. website which contains two csv files.

GitHub Link: https://github.com/prabhkang1/CIND820-2022

Introduction

In current times, all aspects of daily life are transitioning to digital world which includes cashless transaction activities. The rise of internet has increased the usage of credit cards. Credit cards have become one of the most popular modes of payment for electronic transactions. However, as the number of credit card users are increasing exponentially, so are the credit card frauds and defaulters. Financial Institutions evaluate credit risk based on a customer's credit history. This historical information is analysed to avoid any financial losses to the institution.

The correct assessment for credit card approval is very important for financial institutions who provide credit card to the customers. Along with this, an automatic process is required to fasten the approval or rejection decision of the banks. A wide range of machine learning techniques have been developed to solve credit card related problems.

This capstone project would inspect the dataset taken from Kaggle website which merges the personal information data from the customer and the personal behavior information data. The objective of this project is to identify the most efficient and best performing models that can be used to predict the approval of credit cards based on the attributes of the credit card application. The focus will be on evaluating and comparing machine learning classification models such as Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines, Decision Tree, Random Forest and XGBoost classifier. The performance of each of these classification models would be evaluated based on several performance evaluation measures such as Confusion Matrix, Precision, Recall, F-1 Score, Accuracy and AUC & ROC Curve.

Literature Review

In the past, work has been done on similar research problem and dataset, various machine learning models have been proposed to determine and evaluate the credit scoring criteria. In this project, I collected and analyzed number of research papers published. Authors of these research papers applied various approaches to their research problems. They have applied different machine learning algorithms and compared the accuracy of the models to identify the most effective one.

K.S. Naik (2021) in the research paper builds a credit scoring model to forecast credit defaults for unsecured lending (credit cards), by employing machine learning algorithms. They have applied Synthetic Minority Oversampling Technique (SMOTE), to stabilize the imbalanced data which could cause a challenge in their predictive models. They have applied 7 different classifiers including Logistic Regression, SVM, KNN, Decision Tree, Random Forest, XGBoost and LGBM to the processed data set, they found out that Light Gradient Boosting Machine (LGBM) classifier model is efficient to manage larger data volumes.

Ji-Hui MUN, Sang Woo JUNG (2021) analyzed and predicted the delinquency and delinquency periods of credit loans according to gender, own car, property, number of children, education level, marital status and employment status. They have applied the Linear Regression analysis and enhanced decision tree algorithm in this paper. Their research predicted that the Boosted Decision Tree Algorithm made more accurate prediction.

D.Jayanthi (2018) analysed credit card approval data set from UCI machine learning repository. This is a smaller data set than the data set I have selected for this project. They have proposed a credit scoring model of consumer loans based on various analytical models. They have applied and compared Logistic regression and Classification and Regression tree models. Their research concluded with CART model to be more effective than Logistic regression model.

Siddhi Bansal, Tushar Punjabi (2021) has compared different Supervised Machine learning models in their research paper in order to predict how likely a credit card request would be approved on the basis of the parameters like Precision, Recall, Time, Accuracy, F1-Score. The result of their research indicated that the Random Forest Classifier is the best-suited model according to the F1-Score.

Arokiaraj Christian St Hubert, R. Vimalesh, M. Ranjith, S. Aravind Raj (2020) in their work has attempted to improve the available technology using decision tree algorithm, K Nearest Neighbor algorithm and Logistic Regression algorithm. Authors performed data collection, data cleaning, data analysis and visualization, and data splitting tasks before applying machine learning models to the processed dataset. They processed the trained and tested data through the above-mentioned algorithms to get the best accuracy result. They concluded that both Decision Tree and KNN algorithms provided good results after continuous training of different sets of collected data.

Md. Golam Kibria and Mehmet Sevkli (2021) have built a deep learning model which could support the credit card approval decision. Then, they compared the performance of their model

with other two traditional machine learning algorithms, Logistic Regression, and Support Vector Machine. They have pre-processed and analysed the data and applied grid search technique to find the best parameters. Their result show that the overall performance of the deep learning model was slightly better than the other two models.

Therefore, after reading the papers, I observe that solving this problem has many approaches, and every model applied would lead to prediction with different accuracies.

Data Description

Data Source: The data set I have obtained is from Kaggle Inc. website. There are two csv files: application_record.csv and credit_record.csv.

Data Description: To predict the credit card approval decision, I would be using and combining two sets of data, one would contain the customer's personal information and other contains the customer's behaviour patterns.

Application_record.csv consists of personal information of the customers. It contains total of 18 data including ID, gender, car, number of children, total income, education level, family status, housing type, birth date, employment, phone, email, occupation, and family size. There are total of 18 columns and 438,557 rows.

Feature Name	Feature Content
ID	Customer number
CODE_GENDER	Gender
FLAG_OWN_CAR	Car ownership
FLAG_OWN_REALTY	Property ownership
CNT_CHILDREN	Number of children
AMT_INCOME_TOTAL	Total Income
NAME_INCOME_TYPE	Income type
NAME_EDUCATION_TYPE	Education level
NAME_FAMILY_STATUS	Marital status
NAME_HOUSING_TYPE	Housing status
DAYS_BIRTH	Birth date
DAYS_EMPLOYED	Employment start date
FLAG_MOBIL	Mobile phone
FLAG_WORK_PHONE	Work phone
FLAG_PHONE	Phone
FLAG_EMAIL	Email
OCCUPATION_TYPE	Occupation type
CNT_FAM_MEMBERS	Family size

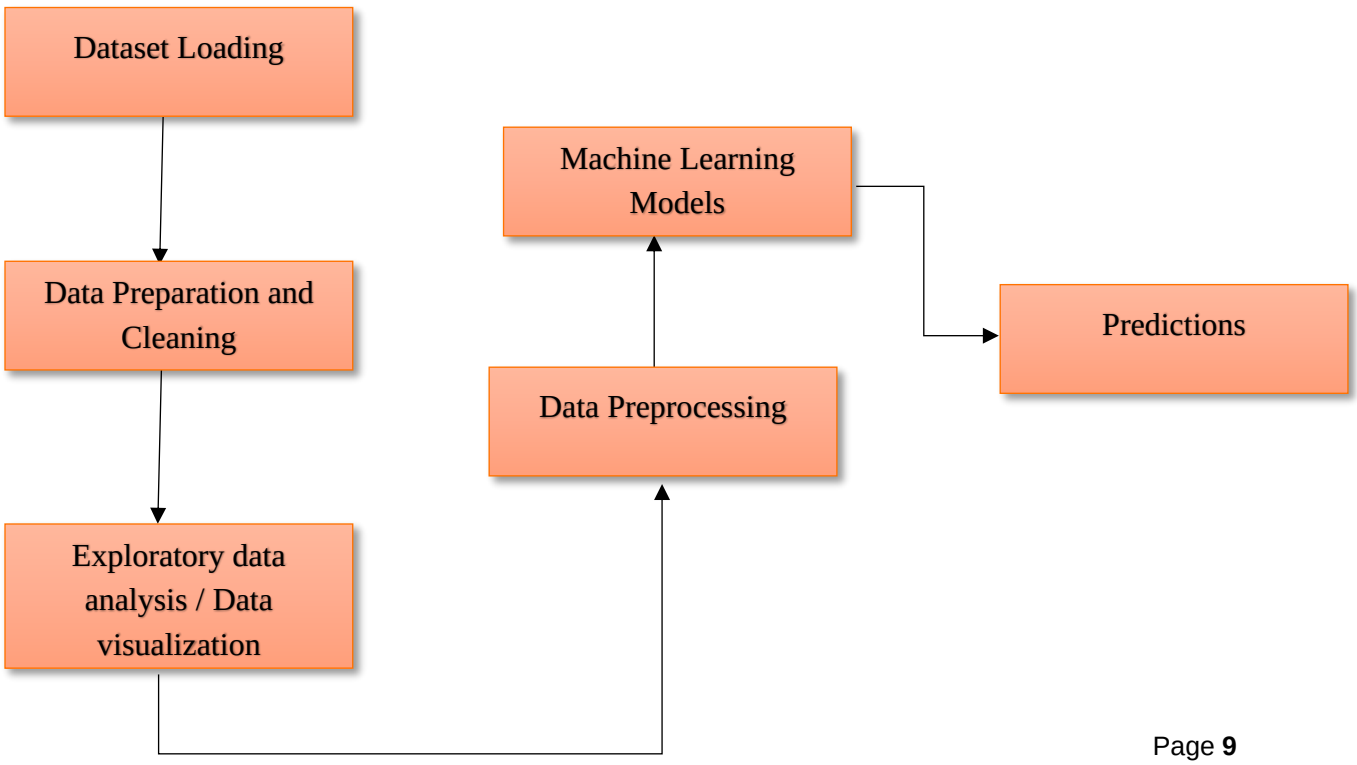
Table 1: Application_record.csv data description

Credit_record.csv is a dataset that contains credit card user's behavior pattern. Data consist of ID, Monthly balance, and status of the credit card user. MONTH_BALANCE variable proceeds in reverse order, the month of extracted data is the starting point. O is the current month, -1 is the previous month, and so on. There are total of 3 columns and 1,048,575 rows

Feature Name	**Feature Content**
ID	Customer number
MONTHS_BALANCE	Record month
STATUS	Status of monthly payment

Table 2: Credit_record.csv data description

Approach Overview



References

1. Dataset: Kaggle Inc. Website, Available at: https://www.kaggle.com/rikdifos/credit-card-approval-prediction

2. Naik, K. S. (2021). Predicting Credit Risk for Unsecured Lending: A Machine Learning Approach. *arXiv preprint arXiv:2110.02206*. Available at: https://arxiv.org/pdf/2110.02206.pdf

3. Shin, W. -S., & Shin, D. -H. (2020). A Study on the Application of Artificial Intelligence in Elementary Science Education. *Journal of Korean Elementary Science Education, 39(1), 117-132.* Available at: https://www.koreascience.or.kr/article/JAKO202116758671173.pdf

4. Bansal, Siddhi, & Punjabi, Tushar. (2021). Comparison of Different Supervised Machine Learning Classifiers to Predict Credit Card Approvals. *International Research Journal of Engineering and Technology (IRJET), E-ISSN: 2395-0056, P-ISSN: 2395-0072, 8(3), 1339-1348, March 2021.* Available at: https://www.irjet.net/archives/V8/i3/IRJET-V8I3277.pdf

5. D.Jayanthi. (2018). Credit Approval Data Analysis Using Classification and Regression Models. *IJRAR-International Journal Of Research And Analytical Reviews (IJRAR), E-*

ISSN 2348-1269, P- ISSN 2349-5138, 5(3), 162-169, September 2018. Available at: https://www.ijrar.org/papers/IJRAR190B030.pdf

6. Arokiaraj Christian St Hubert, R.Vimalesh, M. Ranjith, & S. Aravind Raj. (2020). Predicting Credit Card Approval of Customers Through Customer Profiling using Machine Learning. *International Journal of Engineering and Advanced Technology (IJEAT), 9(4), 52-557.* Available at: https://www.ijeat.org/wp-content/uploads/papers/v9i4/D7293049420.pdf

7. Kibria, Md & Şevkli, Mehmet. (2021). Application of Deep Learning for Credit Card Approval: A Comparison with Two Machine Learning Techniques. *International Journal of Machine Learning and Computing, 11(4), 286-290, July 2021.* Available at: 10.18178/ijmlc.2021.11.4.1049.