CREDIT CARD APPROVAL PREDICTION USING PYTHON

ABSTRACT:

This project aims to build a model that can give results on whether a financial institution can approve credit cards to its customer. This card approval decision by financial companies is done based on considering various reasons related to individuals varying from creditworthiness, loan and repayment history, and income standards. This model can help an institution to make a precise judgment on whether a card can be approved or denied for avoiding fraudulence that can impact financial companies with loss. Through the project work, I tried to examine what are the keynote features or requirements considered for issuing a credit card to consumers by financial institutions by evaluating the existing data set from a machine learning repository through machine learning visualization and analysis techniques.

INTRODUCTION:

In Current times, everything is completely changed as a digital attribute. One of those digitalized areas is cashless transaction activity. This is very common nowadays, and more people are inclined towards this as this reduces the risk of misplacing cash physically. So, many financial institutions are providing cashless means for their users like debit and credit cards. One of the most prominent options is a credit card. Most people rely on credit cards to perform their transaction activities as it is a very easy way of making their payments. The decisiveness by many financial institutions like national and private banks rely on consumer information like their basic info, living standards, salary, yearly and monthly returns, their current livelihood income source. All this info is reviewed for considering an application. This complete check and analysis can avoid bearing a lot of technical and non-technical losses to the institution. This proper analysis is required as we see tremendous growth in this business sector to avoid any kind of potential risk related to the unethical consumer. Precise verification needs to be incorporated by banks when granting credit cards to the applicant. Even though decision-making differs from bank to bank,the most common factor considered by financial institutions is the consumer's credit score. As we are seeing an increase in the large growth margin of the credit business of the financial institution due to more consumers interested in applying for credit cards, there is a need to completely automate the process in order to fasten the approval decision by banks. This helps the bank in improving business along with saving time and need of less manpower which is a major saving in terms of money. The model needs to identify the consumers who applied for credit card into two sectors: "No Risk Present" which means the bank can lend money and there is a guarantee that consumer will pay back and banks will not undergo any risk and loss and "risk present" which means banks shouldn't approve any credit because there is a high chance that consumer can do fraud and banks can undergo financial loss. This classification is done by considering various factors of the

consumer like age, salary, the number of years he/she has been working, yearly income, assets, source of income, credit score, repay history, and existing loan dues. These entire mechanisms are not only applicable for a single consumer, but also to business whether large scale or small scale. In the past, there were various methods introduced to examine the loan history of the consumer and to improve the precision of credit score (Banasik and Crook 2010 and Thomas et al 2002). These models are data mining models that can be categorized into data that depends on statistical distribution and data which does not necessarily depend on the distribution of data. The best example of the model which relies on data distribution is the logistic regression and linear regression analysis model. The linear regression model analysis is used in generating credit scores but this analysis is not favorable since data considered for approving and declining are completely different. Logistic regression supports the data unlike the linear model for parametric tests. Other decision support tools like decision trees , vector machine support are used for non-parametric tests in machine learning. In recent research going on data mining, hybrid approach-based methods are giving optimal results. The neural network approach is considered a better approach to increasing the accuracy of the credit score prediction This paper proposes a model that predicts whether a credit card can be issued or declined to the applicant by a financial institution. Even though the decision of the banks is unique and made based on their organization-designed rules, there are certain similar features that are considered similar and those features are taken into consideration when the algorithm is implemented. Data is taken from a publicly available machine learning repository, the University of California repository. Machine learning preprocessing techniques are implemented and data transformation techniques like scaling, handling of missing values by predefined methods called mean imputation and label encoding for numeric and non-numeric data, dividing data into test and train sets, applying classifiers and to conclude the paper obtained results are further examined using metrics like confusion matrix to examine the accuracy of the result.

ARCHITECTURE:

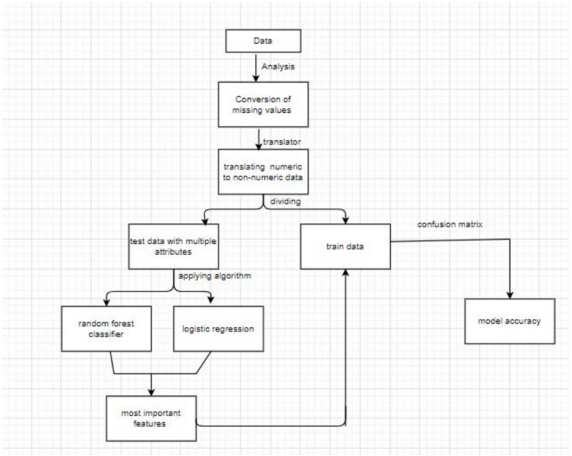


Fig 1: Architecture of Algorithm

The implementation of the project is done in multiple steps by applying various techniques. The steps vary from analysis of dataset by observing, processing the data by identifying anomalies or data that is needed

to be converted since dataset available can be masked for various security reasons. Further, handling the missing values in the dataset taken, then dividing data into two sets such that one set is used for training so that we can develop the model and another set is used for testing and verifying the model for accuracy.

IMPLEMENTATION:

The first step is to obtain the dataset. The dataset used for this paper is downloaded from the publicly available University of California, Irvine machine learning repository. This repository contains a total of 690 records of the consumers of the bank. These records are a mix of both approved and non-approved credit card applications. A glance can tell that complete records are altered into information that cannot be interpreted since data is confidential. Further, to protect the 10 privacy of information, data that can be seen is encrypted into an unrelated format. To understand data, loading and inspecting data methods are imposed on the dataset. The conclusion is that the dataset contains alphanumeric values, and some attribute row values are empty or absent. To perform data visualization analysis, a software library of python called pandas is used.

The dataset consists of alphanumeric data. Age, debt, years employed and credit score columns in the dataset contain numeric values (int and float) and the rest of the columns are object type. The dataset contains 590 rows and 16 columns of data. There are some existing missing values that cannot be ignored because they can affect the accuracy of the model. We overwrite question marks with NAN by implementing the replace method from the NumPy library. The method called “mean imputation” is performed on columns with only numerical data and replaces NAN with the mean value. Mean imputation is applied over the dataset so that data can be converted or replaced with existing variables' mean value. If the row values are numeric, the mean imputation mechanism is implemented and if the row values are non-numeric (object type) then the missing column value will be replaced with the most frequent occurred value in that column of the dataset.The exploratory data analysis approach is opted for categorizing features of numeric values to visualize the actual structure of data. For translating the non-numeric data, we convert data into numeric by using Label Encoder class and convert all the object types to numeric by enforcing the Label Encoder method for conversion. Figure 6 shows the conversion of non-numeric data into numeric data.

Before incorporating the model on the dataset, the dataset is split into two subsets: a training set and a test set. The training dataset is for developing the model and the test set is for checking the accuracy of the developed model. Seventy-five percent of the data in the dataset is a training set and the remaining twenty-five percent of the data is a testing set. Since the dataset has varying ranges, scaling is applied to training data to transform their range to be between 0 to 1. Approval or acceptance of the card is a

fixed decision like it could be either yes or no. Therefore, we can consider the output to be a binary attribute. So, the binary classes model well-known in data science is logistic regression. Logistic regression is an algorithm that uses functions to determine the occurrence output. it uses the mechanism called “the most likely chance of the event occurring percentage.” Random forest classifier is a machine learning algorithm used for both classification and regression. It randomly takes samples from the dataset and constructs a decision tree for generating the prediction. Using this algorithm, we can identify the features that play a crucial role based on 14 the relevance score for making predictions by using the feature variable to view the score and visualize the score via graph representation using the seaborn library

DATASET & METHODOLOGY:

Dataset Description

We utilize the Credit Card Approval dataset from the UCI Machine Learning Repository. This dataset comprises diverse features of applicants, including demographic details (age, gender, education), financial information (income, credit score, employment status), and application specifics (amount requested, type of card). The target variable indicates approval status (approved/rejected).

Data Preprocessing

Prior to model training, we carefully preprocess the data to ensure its suitability for machine learning algorithms. This includes handling missing values through imputation or deletion based on their characteristics and impact. We then encode categorical features using one-hot encoding to transform them into numerical representations understandable by the model. Finally, we standardize or normalize numerical features to ensure consistent scales and avoid bias towards features with larger values.

Model Training and Evaluation

We implement a Logistic Regression model using the scikit-learn library in Python. To train and evaluate the model effectively, we split the dataset into training and testing sets. We utilize GridSearchCV or RandomizedSearchCV for hyperparameter tuning, optimizing the model's performance by adjusting parameters such as the regularization constant (C). To assess the model's effectiveness, we employ various evaluation metrics, including accuracy, precision, recall, F1-score, and the AUC-ROC curve.

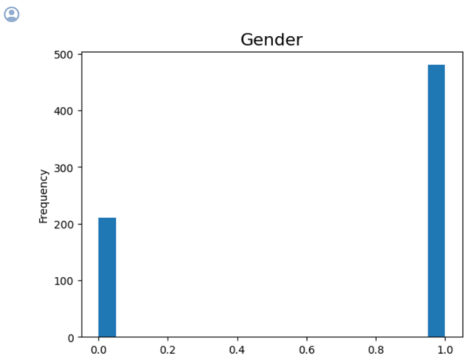
IMPLEMENTATION AND RESULTS

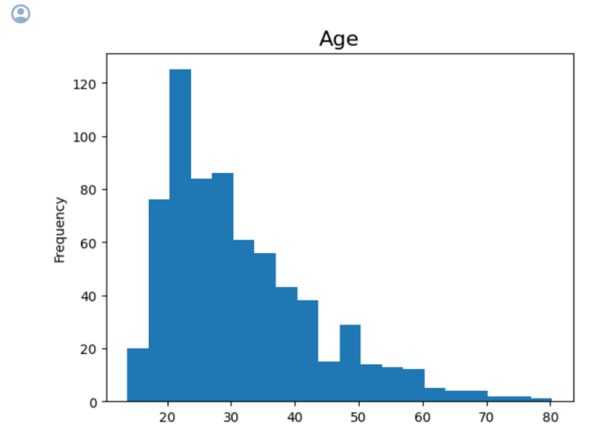
This chapter describes implementation and results evaluation activities in detail. List of activities discussed in here are explanatory analysis of data, data preparation activities,

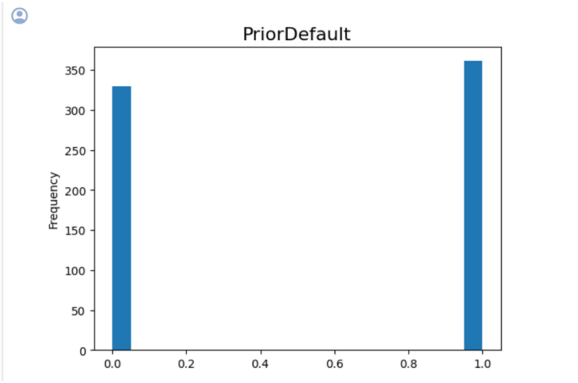
models building, evaluation of models and deployments.

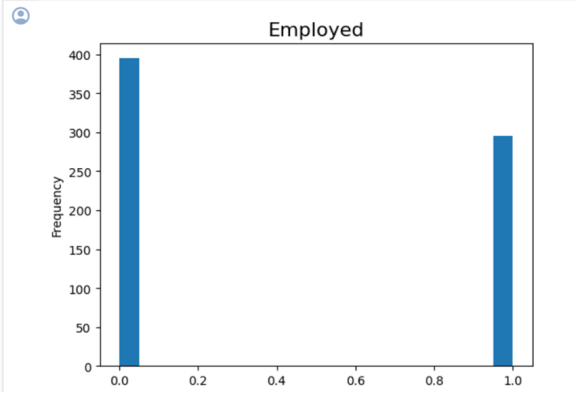
Explanatory Data Analysis

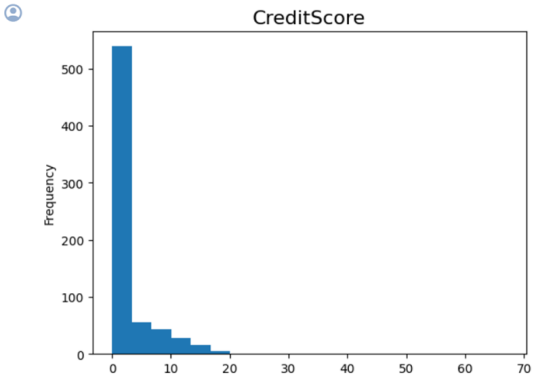
- Graphical and numerical representation of data provide better insight about particular data set.
- Graphical representation of our data set is described below.

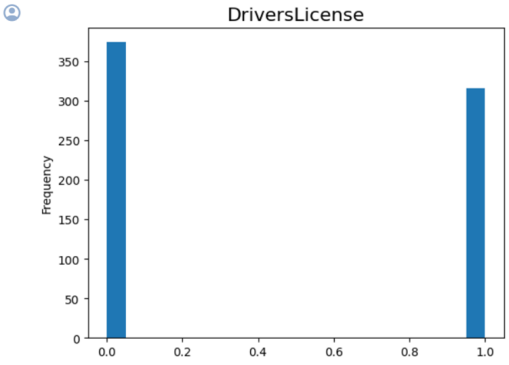




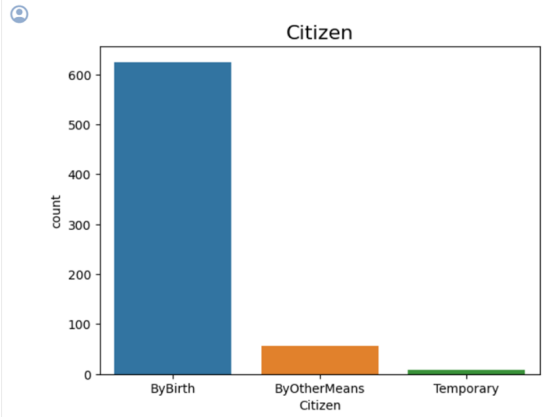


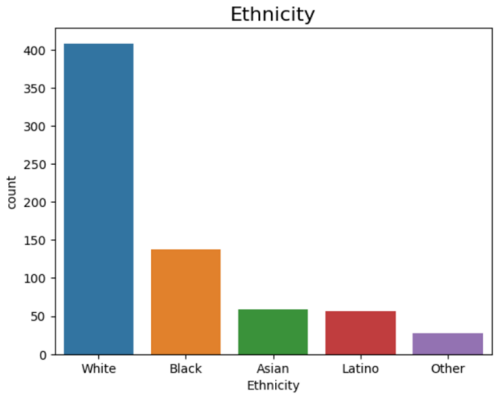


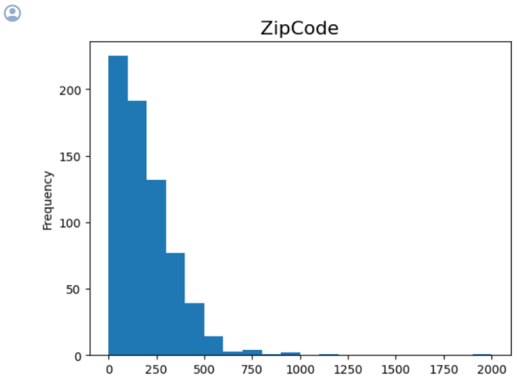


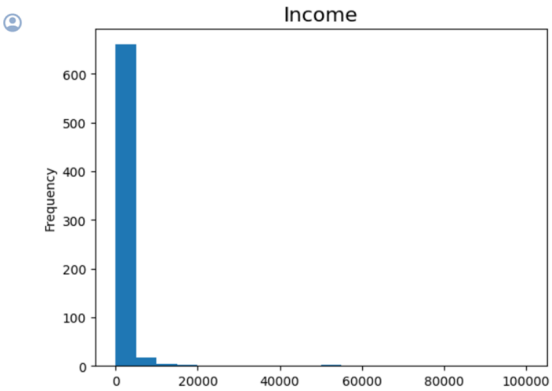


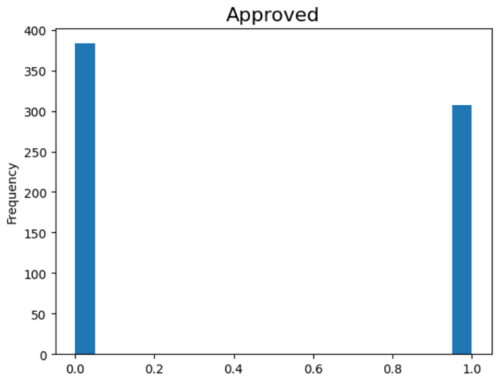
CREDIT CARD APPROVAL PREDICTION USING PYTHON











ERRORANALYSIS: UNDERSTANDING PRECISION AND RECALL

A comprehensive and insightful error analysis is crucial for understanding the model's limitations and identifying areas for improvement. We analyze the confusion matrix to determine the number of True Positives (correctly predicted approvals), True Negatives (correctly predicted rejections), False Positives (incorrectly approved applications – risky borrowers), and False Negatives (incorrectly rejected applications – worthy borrowers).

MODEL IMPROVEMENT STRATEGIES: BEYOND LOGISTIC REGRESSION

Based on the error analysis and understanding the business context, we propose several strategies to improve model performance:

- Refined Hyperparameter Tuning: Analyze precision and recall

performance across different parameter settings to optimize for the desired trade-off.
- Feature Engineering: Craft new features that capture additional relevant information from the data, potentially leading to improved accuracy.
- Ensembling Techniques: Combine Logistic Regression with other models like Random Forest or Gradient Boosting to leverage their strengths and potentially achieve higher accuracy and generalizability.
- Deep Learning Approaches: Explore advanced architectures like deep neural networks for potentially improved performance, especially when dealing with complex data relationships.

RESULT:

- Most pronunced difference is when age is higher. 40 and above
- Also much more of the errors we are concerned about had shown prior defaults.
- Slightly more genders as 0**.**

CONCLUSION:

During the development of our credit card approval predictor, we honed our expertise in various preprocessing techniques essential for refining raw data into a model-ready state. This involved meticulous steps like feature scaling, label encoding for categorical variables, and adeptly handling missing data, ensuring our models were fed clean, standardized inputs for optimal performance.

In our pursuit of the most accurate prediction, we didn't settle for just one model. Instead, we curated a repertoire of five distinct machine learning models, fine-tuned their hyperparameters with precision, and meticulously evaluated their performances. Our evaluation wasn't limited to mere accuracy scores; we delved deep into cross-comparisons between the training and test datasets, ensuring our models weren't just memorizing but understanding the underlying patterns.

Harnessing the power of Python's robust machine learning libraries, we orchestrated an ensemble of algorithms, leveraging the versatility and efficiency of tools like scikit-learn and TensorFlow. This allowed us to craft a predictive framework that not only forecasts credit card approval but does so with a refined, accurate, and reliable methodology.

In essence, our journey wasn't just about building a predictive model; it was a testament to our dedication to precision, optimization, and the relentless pursuit of refining the art and science of machine learning for real-world applications like credit card approval prediction.

CREDIT CARD APPROVAL PREDICTION USING PYTHON

REFERENCES:

1. Koh, H. C., & Chan, K. L. G. (2002). Data mining and customer relationship marketing in the banking industry. Singapore Management Review, 24(2), 1–27.
2. S. B. Kotsiantis. Supervised Machine Learning: A Review of Classification Techniques. Informatica 31 (2007) 249-268 Web
3. Dataset, https://raw.githubusercontent.com/brandynewanek/brandynewanek/main/creditcard_clean_dataset.csv
4. http://www.cs.stir.ac.uk/courses/ITNP4B/lectures/kms/4-MLP.pdf
5. http://localhost:8888/notebooks/Downloads/MajorProjectBM/Predicting%20Credit%20Card%20Approvals/notebook.ipynb
6. Medium. 2021. Credit Card Approval Prediction Model in Python. [online] Available at: < https://medium.com/@ashish.tripathi1207/credit-card-approval-prediction-model-in-python-c0e07677058e > [Accessed 29 June 2021].
7. http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html
8. Http://rstudio-pubs-static.s3.amazonaws.com/73039_9946de135c0a49daa7a0a9eda4a67a72.h**tml**
9. 1.1. Linear Models — scikit-learn 1.2.2 documentation.” [Online]. Available: https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
10. Credit Card Approvals (Clean Data).” [Online]. Available: https://www.kaggle.com/datasets/samuelcortinhas/credit-card-approval-clean-data
11. Ensemble learning,” Apr. 2023, page Version ID: 1151030544. [Online].Available: https://en.wikipedia.org/w/index.php?title=Ensemble_learning&oldid=1151030544
12. Visa Inc.” Apr. 2023, page Version ID: 1151894415. [Online]. Available:https://en.wikipedia.org/w/index.php?title=Visa_Inc.&oldid=1151894415
13. G. Bertola, R. Disney, and C. B. Grant, The Economics of Consumer Credit.MIT Press, 2006, google-Books-ID: X8ZaBJpRiTsC.

Presented by

Tejaswini Paritala
tp77290n@pace.edu

YashaswiniVardhamanukota
yv62330n@pace.edu