Credit Card Approval System

Vineet Veeresh Kiragi

San Jose State University

Problem Definition:

The problem definition is the statement that clearly outlines the specific issue or challenge that needs to be addressed. In the context of the project " Credit Card Approval Predictions Using Logistic Regression," the problem definition can be stated as follows:

The objective of this project is to develop a credit card approval system using Logistic Regression, Decision Tree Classifier, and Random Forest Classifier. The system aims to predict whether a credit card application should be approved or rejected based on various customer attributes. In today's fast-paced world, where financial transactions are increasingly digital and convenient, the need for an efficient and accurate credit card approval system has become crucial. Credit card approval systems play a vital role in the financial industry by enabling banks and credit card companies to make informed decisions about whether to grant credit to individuals. These systems streamline the application process and reduce the time and effort required for manual evaluation of each application. By leveraging machine learning algorithms, such as Logistic Regression, Decision Tree Classifier, and Random Forest Classifier, credit card approval systems can analyze large volumes of data and provide fast and reliable credit decisions.

By utilizing machine learning algorithms, the credit card approval system can consider multiple factors simultaneously, including gender, employment status, income, education level, and other attributes from the dataset. These algorithms can identify complex patterns and relationships within the data, allowing for accurate credit risk assessment and efficient decision-making.

Project Objectives:

The project objectives are the specific goals and outcomes that the project aims to achieve. In the context of the project " Credit Card Approval Predictions Using Logistic Regression," the following objectives can be defined:

* **Develop a Credit Card Approval System:** The primary objective of this project is to develop a robust and accurate credit card approval system using three different machine learning algorithms: Logistic Regression, Decision Tree Classifier, and Random Forest Classifier. The system will utilize a dataset containing various customer attributes such as gender, car ownership, property ownership, employment status, family size, income, education level, and more. By training these algorithms on historical data, the system will learn patterns and relationships that can predict credit card approval decisions with a high level of accuracy.
* **Create a User-Friendly Front-End Application:** In addition to the credit card approval system, another objective is to develop a user-friendly front-end application that allows users, such as bank employees or credit card applicants, to interact with the system seamlessly. The front-end application will be developed using HTML, CSS, and JavaScript, providing an intuitive and visually appealing interface. Users will be able to input their personal information and receive real-time credit card approval predictions based on the trained models.
* **Host the System on Flask:** To ensure the credit card approval system is accessible and can handle user requests effectively, the project aims to host the system on Flask, a popular Python web framework. Flask provides a lightweight and efficient server environment for hosting web applications. By leveraging Flask, the system will be able to handle multiple user requests simultaneously, ensuring a smooth and responsive user experience.
* **Ensure System Reliability and Security:** Another important objective of this project is to ensure the reliability and security of the credit card approval system. Reliability is crucial to handle high volumes of user requests without system failures or slowdowns. The system will be designed with robust error handling and exception management to provide a seamless user experience. Additionally, security measures, such as encryption of sensitive data and implementing authentication mechanisms, will be implemented to protect user information and prevent unauthorized access.
* **Ensure ethical considerations in data collection and usage**: This objective focuses on incorporating ethical practices throughout the project. It includes ensuring the privacy and confidentiality of individuals' data, avoiding biases in data collection and model development, and considering the potential impacts on individuals and society. The project should adhere to best practices in data ethics and responsible data stewardship.
* **Provide documentation and knowledge transfer**: This objective involves documenting the project's processes, methodologies, algorithms, and findings. It aims to provide comprehensive documentation that enables replication, further research, and knowledge transfer to other stakeholders or organizations interested in credit card fraud detection.

Analysis:

**Data Exploration:**

Before proceeding with model training, a thorough analysis of the dataset was conducted to gain insights into the data distribution, identify any data inconsistencies or outliers, and understand the relationships between different attributes. The analysis involved descriptive statistics, data visualization techniques, and data profiling.

Data visualization techniques such as histograms are employed to visualize the distribution and relationships between variables. These visualizations will aid in identifying potential patterns, outliers, or correlations within the data. The identified outliers are then eliminated using outlier removal techniques.

Descriptive statistics provided summary statistics such as mean, median, standard deviation, and quartiles for numeric variables, while frequency counts will be obtained for categorical variables. This helped in understanding the central tendency, spread, and distribution of the data.

Data profiling involved examining the completeness and quality of the dataset. The dataset is checked for missing values, duplicate records, inconsistent data formats, and outliers. Missing values were handled through removal, depending on the extent of missingness and the nature of the data.

**Data Preprocessing:**

Data preprocessing is a crucial step to ensure the quality and suitability of the data for model training. Several preprocessing techniques were applied to prepare the dataset for the machine learning algorithms.

Categorical variables were encoded using techniques such as label encoding to convert them into numerical representations. This transformation is necessary for the machine learning algorithms to interpret and utilize the categorical information effectively.

Numeric variables may require feature scaling to bring them to a common scale and prevent bias during model training. Common scaling techniques include standardization (mean centering and scaling by standard deviation) or normalization (scaling to a range of 0-1). Normalization scaling is used to normalize the numerical variables.

The target variables were not equally distributed throughout the 2 classes. SMOTE was used to overpopulate the target values with less no of entries.

Additionally, feature selection or dimensionality reduction techniques such as PCA were applied to eliminate irrelevant or redundant features that do not contribute significantly to the prediction task by retaining 95% of the information. This can improve model efficiency and mitigate the curse of dimensionality.

**Model Training and Evaluation:**

Three different machine learning algorithms will be implemented and evaluated for credit card approval prediction: Logistic Regression, Decision Tree Classifier, and Random Forest Classifier.

Model training will involve splitting the preprocessed dataset into training and testing subsets of 80% of the entries for training and 20% for testing. The training set was used to train the models, while the testing set was used to evaluate their performance. The models will learn from the patterns and relationships in the training data, enabling them to make predictions on unseen data.

To assess the performance of each model, evaluation metrics such as accuracy, precision, recall, and F1-score were calculated. Accuracy, which measures the overall correctness of the predictions, will be the primary evaluation metric. However, considering the nature of the credit card approval task, other metrics such as precision (proportion of correctly predicted positive instances), recall (proportion of actual positive instances correctly predicted), and F1-score (harmonic mean of precision and recall) were also be analyzed to evaluate the models comprehensively.

Cross-validation techniques, such as k-fold cross-validation, may be employed to assess the models' performance across multiple iterations and ensure the results are robust and unbiased.

**Hyperparameter Tuning**

To further optimize the model performance, hyperparameter tuning techniques were employed. Hyperparameters are configuration settings that are not learned from the data and can significantly impact the model's performance. Grid search CV is used to explore different combinations of hyperparameters and identify the optimal settings that maximize the model's performance metrics.

By conducting a thorough analysis of the dataset, applying appropriate preprocessing techniques, training, and evaluating multiple models, and tuning the hyperparameters, the credit card approval system can achieve a high level of accuracy and reliability. The analysis phase plays a crucial role.

Results:

After conducting the analysis and model training, the credit card approval system using Logistic Regression, Decision Tree Classifier, and Random Forest Classifier produced the following results:

|  |  |
| --- | --- |
| **Algorithm** | **Accuracy** |
| Logistic regression | 85% |
| Decision tree | 86% |
| Random forest | 85% |

The performance of each model was evaluated using various evaluation metrics such as accuracy, precision, recall, and F1-score. The accuracy measures the overall correctness of the predictions, while precision focuses on the proportion of correctly predicted positive instances (approved credit cards). Recall measures the proportion of actual positive instances correctly predicted, and F1-score provides a balanced measure between precision and recall.

Overall, the results obtained from the credit card approval system provide a comprehensive understanding of the models' performance, feature importance, real-world application, and performance comparison. These results demonstrate the effectiveness and reliability of the system in predicting credit card approvals and assist financial institutions in making informed decisions while minimizing risks.

Discussion:

The discussion section aims to analyze and interpret the results obtained from the credit card approval system using Logistic Regression, Decision Tree Classifier, and Random Forest Classifier. It provides insights into the system's performance, limitations, and potential implications for real-world credit card approval processes.