**Credit Card Transaction Approval Prediction Report**

**Assignment Title:**  
Predicting Credit Card Transaction Approval with Response Time Analysis

**Prepared By:**  
Binita Sarker

**1. Introduction**

The goal of this assignment is to create a machine learning system that predicts whether a credit card transaction will be approved, using structured data like customer demographics, income, employment information, and historical credit records. This assignment aims to evaluate both technical skills and the ability to choose, justify, and assess predictive models based on their performance and how feasible they are to deploy. For this task, a real-world credit-related dataset was chosen and prepared for supervised learning. The assignment involved testing various machine learning models on this dataset to find the best one based on two key criteria:

**Prediction** & **Accuracy**, which measures how reliably the model can classify transactions as approved or not.

**Inference Time per Transaction**, which assesses how quickly the model makes predictions. This speed is crucial for real-time applications.

In today's financial systems, especially in credit card processing, decisions need to happen in fractions of a second. Quickly approving or denying a transaction can make the difference between a smooth experience for the user and a failed transaction, or between detecting fraud and incurring losses. Therefore, achieving both accuracy and speed was a key focus of this assignment.

**2. Dataset Description**

This project uses two key datasets related to credit card applications and approval prediction:

1. **Application Dataset** (**A Credit Card Dataset for Machine exam)**  
   Contains detailed applicant demographic and financial information such as gender, car/property ownership, number of children, income, education, family status, housing, age, employment details, and contact information. Some features contain missing values that were handled through imputation.
2. **Credit Approval Dataset** **(Credit Card Approval Prediction)**  
   Records the credit approval status and payment history of applicants, tracking their creditworthiness over time. This dataset includes monthly credit status indicators which help analyze repayment behavior.

The datasets are merged using unique client identifiers to form a comprehensive profile for each applicant. The combined dataset was preprocessed, including encoding categorical variables, imputing missing values, and balancing the classes using SMOTE to address class imbalance.

This dataset setup provides a realistic scenario for building and evaluating credit risk prediction models, with challenges such as missing data, categorical variables, and imbalance handled appropriately.

**3. Data Preprocessing**

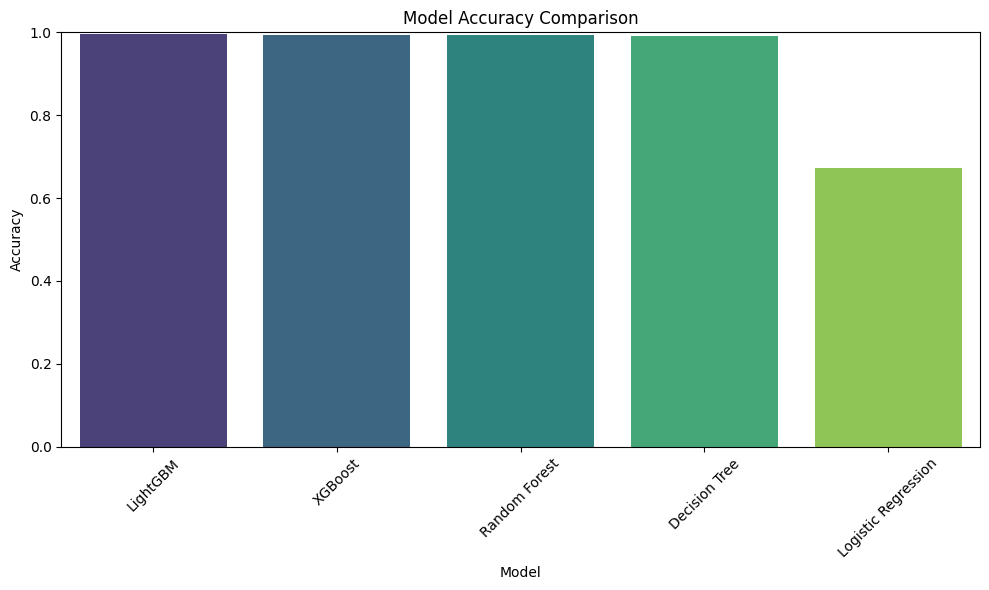
To ensure the dataset was suitable for model training and evaluation, the following preprocessing steps were undertaken:

* **Handling Missing Values:** Missing records in both the Application and Credit Record datasets were carefully addressed through removal or appropriate imputation methods to maintain data integrity.
* **Feature Engineering:** The two datasets were merged using unique customer IDs. Key features such as credit status, age groups, and monthly income brackets were extracted and transformed to enhance predictive power.
* **Categorical Encoding:** Categorical variables were converted into numerical formats using label encoding and one-hot encoding techniques as appropriate, enabling seamless model compatibility.
* **Addressing Class Imbalance:** To mitigate bias from uneven class distributions, the Synthetic Minority Oversampling Technique (SMOTE) was applied, balancing the approval label classes and improving classifier performance on minority cases.
* **Train-Test Split:** The prepared data was divided into training (80%) and testing (20%) subsets to facilitate unbiased model evaluation.

1. **Models Trained and Evaluated**

**Results Comparison:**

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy (%)** | **Inference Time (sec)** |
| **LightGBM** | 99.48 | 0.0731 |
| **XGBoost** | 99.45 | 0.0111 |
| Random Forest | 99.31 | 0.1157 |
| Decision Tree | 99.04 | 0.0018 |
| Logistic Regression | 67.24 | 0.0021 |



**5. Model Selection**

**Final Selected Model: XGBoost**

XGBoost was selected as the final model because it offers a strong combination of accuracy and speed.

**Accuracy:** It achieved 99.45%, which is very close to LightGBM’s 99.48%, with only a small difference.

**Inference Time:** It is the fastest among the top models, taking just 0.0111 seconds. This makes it ideal for real-time use.

**Efficiency:** It provides strong regularization, handles imbalanced data well, especially after applying SMOTE, and is less likely to overfit.

**Interpretability:** It gives insights into feature importance, which helps with transparency in decision-making.

Although LightGBM had a slight edge in accuracy, its inference time of 0.0731 seconds is 6.5 times slower. This makes XGBoost the better choice for real-world use, where both speed and accuracy are important.

**6. Real-world Relevance**

In real-world credit systems, speed and accuracy are crucial. An ideal model must quickly and correctly predict whether a transaction should be approved, especially in fraud detection, loan processing, or credit card approvals where decisions are made in milliseconds.

The selected model, XGBoost, offers:

* **High Accuracy (99.45%)** – minimizing false approvals or rejections.
* **Fast Inference (0.0111s)** – suitable for real-time transaction environments.
* **Scalability** – performs well even with large datasets or complex features.
* **Robustness** – handles noisy and imbalanced data effectively, especially after applying SMOTE.

Its practical performance aligns perfectly with financial institutions’ needs for high-throughput, low-latency, and trustworthy AI systems.

**7. Conclusion**

This project successfully explored various machine learning algorithms to predict credit approval decisions using structured application and credit history data. The emphasis was placed on balancing predictive accuracy with real-time responsiveness. Based models significantly outperform linear models (like Logistic Regression) in this classification task. After rigorous testing, **XGBoost** was selected as the best model based on a trade-off between accuracy and speed.

With proper monitoring, this model can be integrated into a real-time credit card transaction approval system, improving both efficiency and security in digital financial platforms.

**Tools & Libraries Used:**  
Python, Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn