# North Carolina State University

# Project Report
# On
# Credit Card Approval Prediction

**Fredrick Livingston, Ph.D.**

EM589 Practical Machine Learning For Engineering Analytics

## Abstract

This comprehensive report presents the findings of a project aimed at predicting credit card approvals using machine learning techniques. The project involved various phases, including data preprocessing, model development, feature selection, hyperparameter optimization, model evaluation, and report preparation. The report summarizes the main methodologies used, experimental results obtained, and key insights gained throughout the project.

## Introduction

Access to credit cards plays a significant role in financial inclusion and economic participation. However, the traditional credit approval process can be time-consuming and subjective. Automated credit approval prediction systems powered by machine learning algorithms offer a faster and more objective alternative. This project seeks to leverage machine learning techniques to develop such a system, providing financial institutions with a reliable tool to assess creditworthiness accurately and efficiently.

## Business Understanding

- Leveraging data-driven predictive models to revolutionize the credit approval process, enhancing efficiency and accuracy.
- The banking sector struggles to accurately evaluate client creditworthiness, often relying on error-prone manual processes.
- We develop accurate predictive models, providing insights to banks, fostering financial inclusion and empowerment.
- Using ML algorithms and historical data, we aim to predict credit card approvals accurately, optimizing decisions for financial institutions and empowering individuals financially.

## Related Work

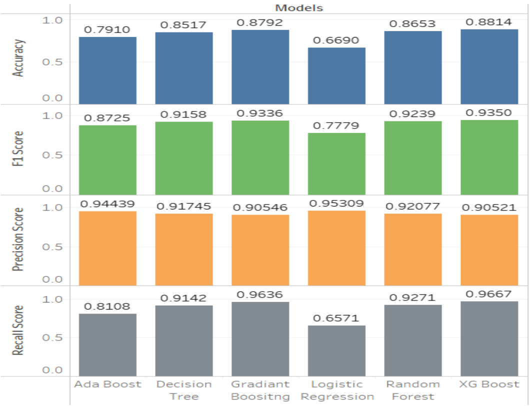
Prior research in credit card approval prediction has explored various machine learning algorithms and methodologies. Similar studies have utilized logistic regression, decision trees, ensemble methods, and neural networks to predict credit card approvals based on applicant information. By reviewing related work, this project builds upon existing knowledge and contributes to the advancement of credit risk assessment techniques.

## Project Roadmap

- **Business Understanding**
  - Exploring the potential of machine learning to expedite credit decisions and foster transparency.
- **Data Understanding**
  - Unveiling hidden patterns in data crucial for precise credit card approval predictions
- **Data Preparation**
  - Creating and refining a clean dataset to improve the accuracy of predictive modelling.

- **Modelling**
    - Customizing machine learning models to suit the unique characteristics of the dataset, optimizing their effectiveness in predicting credit card approvals
- **Evaluation**
    - Conducted comparative analysis to identify and select the most effective model for credit evaluation.
- **Deployment**
    - Integrating the most effective model seamlessly into banking systems to improve credit assessment processes and promote financial inclusion.

## Proposed Methods



The proposed methods for credit card approval prediction include the utilization of several machine learning algorithms, namely logistic regression, random forest, gradient boosting, Adaboost, and XGBoost. These algorithms were chosen for their ability to handle classification tasks and their effectiveness in capturing complex patterns in the data. Additionally, extensive data preprocessing techniques were applied to clean the dataset, handle missing values, and encode categorical variables to prepare it for model training.

Some algorithms used in the modeling are:

### Random Forest Classifier (RF)

- Parameters: 500 decision trees and a random state of 123
- Feature Selection: Implemented Recursive Feature Elimination (RFE)
- Purpose: complex datasets with high dimensionality

### Gradient Boosting Classifier (GB)

- Parameters: 500 decision trees, learning rate of 0.1, maximum depth of 8
- Feature Selection: Utilized Recursive Feature Elimination (RFE)
- Purpose: sequentially improvement of weak learners, increases accuracy

### XGBoost Classifier (XGB)

- Parameters: 500 decision tree, learning rate of 0.1, maximum depth of 8
- Feature Selection: Applied Recursive Feature Elimination (RFE)
- Purpose: efficiency, scalability, and high performance, complex relationships

### Logistic Regression (LR)

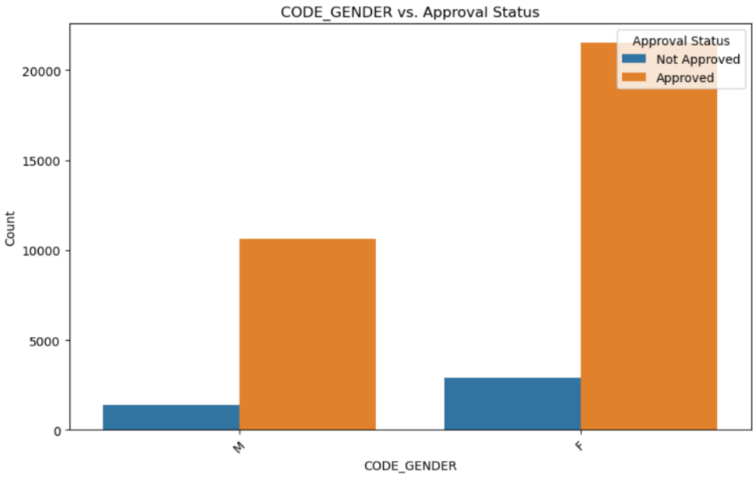
□ Parameters: Default parameters are used

□ Feature Scaling: To ensure the model's convergence and stability

□ Purpose: A baseline model for binary classification tasks, ease of interpretation

□

### Experiments



The experiments conducted during the project involved a systematic approach to model development and evaluation. The dataset used for training and testing the models consisted of various features such as applicant demographics, financial history, and credit card usage patterns. The experimental setup included dividing the dataset into training and testing sets, applying cross-validation techniques, and optimizing model hyperparameters using grid search and randomized search methods. As the data is skewed a technique called SMOTE is used. SMOTE stands for Synthetic Minority Over-sampling Technique. It is used to address the issue of class imbalance, particularly in classification tasks where one class significantly outnumbers the other. Class imbalance can lead to biased models that perform poorly on the minority class.

## Methodology

### Project Initialization

During this initial week, the team focused on project planning, team organization, and dataset acquisition. No significant code updates were made, and no major challenges or blockers were encountered.

### Data Preprocessing

This week focused on data preprocessing and initial exploratory data analysis. The team worked on cleaning the data, handling missing values, and preparing the dataset for model development and feature selection. A data preprocessing pipeline was completed, including handling missing values and encoding categorical variables. Minor delays were encountered due to the complexity of feature engineering.

### Feature Selection and Initial Model Development

This phase marked the beginning of feature selection and initial model development. The team started implementing various machine learning algorithms, such as logistic regression, random forest, decision tree, gradient boosting, AdaBoost, and XGBoost, and evaluated their performance on the pre-processed dataset. Ensuring model interpretability while maintaining high predictive accuracy emerged as a challenge during this week.

### Hyperparameter Optimization

This week primarily focused on hyperparameter optimization and preparing for the upcoming model evaluation. The team delved into fine-tuning the parameters of the machine learning models to enhance their performance. Techniques such as grid search and randomized search were employed for hyperparameter tuning. However, feature selection took longer than expected due to dataset complexity and testing multiple different algorithms.

### Model Evaluation and Selection

During this week, the team conducted comprehensive performance testing on the various machine learning models implemented earlier, analysing their accuracy, precision, recall, and F1-scores on the validation dataset. The team also explored methods to improve model interpretability, ensuring the models could provide meaningful insights. Ensuring a balance between model performance and interpretability emerged as a key challenge.

### Final Model Evaluation

This week was focused on finalizing the model evaluation and selection process, as well as finalizing the model and evaluating it to ensure it was error-free. Code debugging was carried out to ensure there were no logical errors or issues with hyperparameter tuning. The Final Model Evaluation was completed, and work began on the report and final presentation.

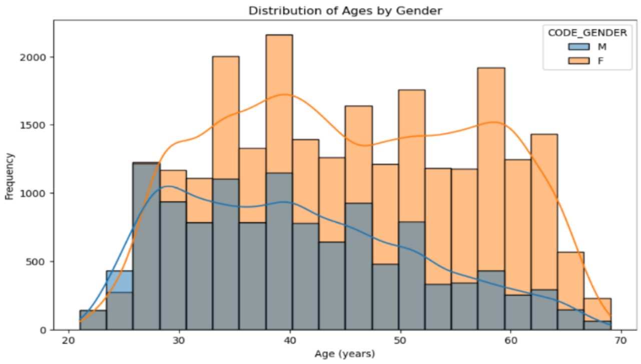
### Final Model Selection

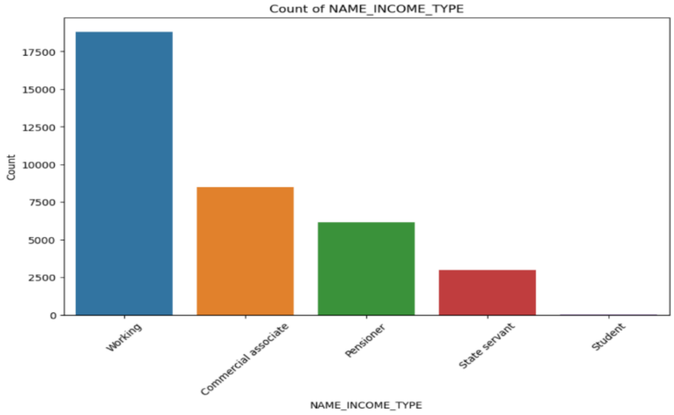
After comprehensive evaluation and analysis, based on the accuracy parameter the team selected the XGBoost Model as the final model for the Credit Card Approval Prediction project. This model demonstrated the best trade-off between predictive performance and interpretability, aligning with the project's objectives and stakeholder requirements.

## Results and Discussion

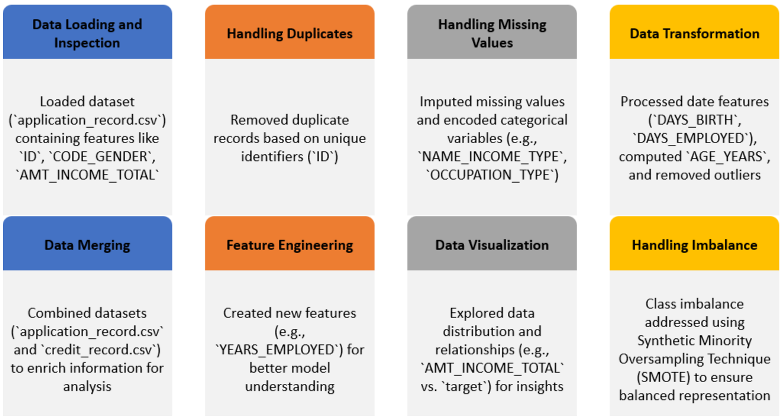
The results obtained from the experiments indicate promising performance of the machine learning models in predicting credit card approvals. The models achieved high accuracy, precision, recall, and F1-scores on the validation dataset, demonstrating their effectiveness in distinguishing between approved and rejected credit card applications. However, challenges were encountered in ensuring model interpretability while maintaining high predictive accuracy, highlighting the need for further research in this area.

## Data Analysis





## Data Preparation



**Data Loading and Inspection**

Loaded dataset (`application_record.csv`) containing features like `ID`, `CODE_GENDER`, `AMT_INCOME_TOTAL`

**Handling Duplicates**

Removed duplicate records based on unique identifiers (`ID`)

**Handling Missing Values**

Imputed missing values and encoded categorical variables (e.g., `NAME_INCOME_TYPE`, `OCCUPATION_TYPE`)

**Data Transformation**

Processed date features (`DAYS_BIRTH`, `DAYS_EMPLOYED`), computed `AGE_YEARS`, and removed outliers

**Data Merging**

Combined datasets (`application_record.csv` and `credit_record.csv`) to enrich information for analysis

**Feature Engineering**

Created new features (e.g., `YEARS_EMPLOYED`) for better model understanding

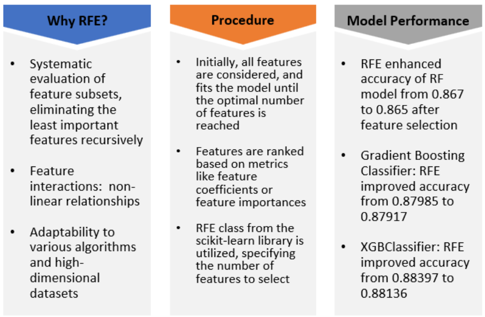
**Data Visualization**

Explored data distribution and relationships (e.g., `AMT_INCOME_TOTAL` vs. `target`) for insights

**Handling Imbalance**

Class imbalance addressed using Synthetic Minority Oversampling Technique (SMOTE) to ensure balanced representation

## Feature Selection



**Why RFE?**

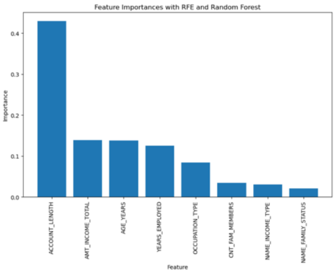
- Systematic evaluation of feature subsets, eliminating the least important features recursively
- Feature interactions: non-linear relationships
- Adaptability to various algorithms and high-dimensional datasets

**Procedure**

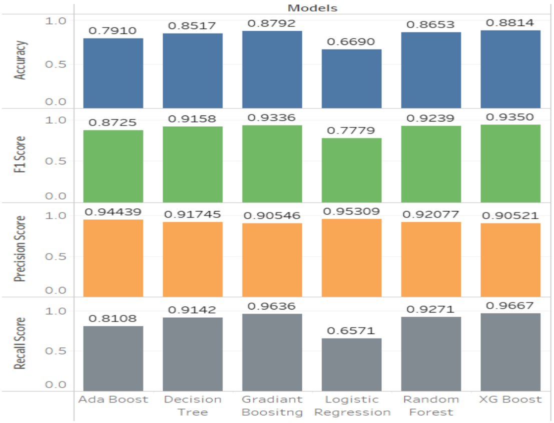
- Initially, all features are considered, and fits the model until the optimal number of features is reached
- Features are ranked based on metrics like feature coefficients or feature importances
- RFE class from the scikit-learn library is utilized, specifying the number of features to select

**Model Performance**

- RFE enhanced accuracy of RF model from 0.867 to 0.865 after feature selection
- Gradient Boosting Classifier: RFE improved accuracy from 0.87985 to 0.87917
- XGBClassifier: RFE improved accuracy from 0.88397 to 0.88136

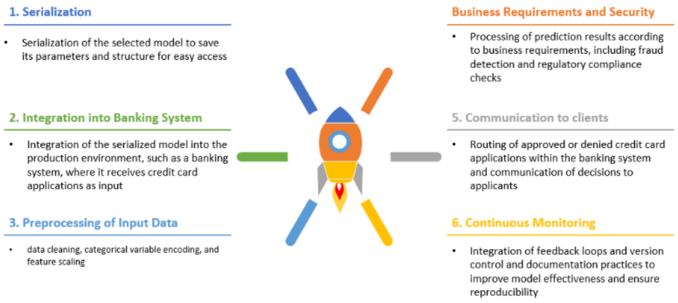


## Model Evaluation



- Confusion matrices were generated to visualize the models' performance in classifying approved and not approved credit card applications.
- The evaluation process involved assessing the models' accuracy scores, F1 scores, precision scores, and recall score.

## Model Deployment



**1. Serialization**

- Serialization of the selected model to save its parameters and structure for easy access

**2. Integration into Banking System**

- Integration of the serialized model into the production environment, such as a banking system, where it receives credit card applications as input

**3. Preprocessing of Input Data**

- data cleaning, categorical variable encoding, and feature scaling

**Business Requirements and Security**

- Processing of prediction results according to business requirements, including fraud detection and regulatory compliance checks

**5. Communication to clients**

- Routing of approved or denied credit card applications within the banking system and communication of decisions to applicants

**6. Continuous Monitoring**

- Integration of feedback loops and version control and documentation practices to improve model effectiveness and ensure reproducibility

The dataset used for training and testing the models consisted of various features such as applicant demographics, financial history, and credit card usage patterns. The experimental setup included dividing the dataset into training and testing sets, applying cross-validation techniques, and optimizing model hyperparameters using grid search and randomized search methods. As the data is skewed a technique called SMOTE is used. SMOTE stands for Synthetic Minority Over-sampling Technique. It is used to address the issue of class imbalance, particularly in classification tasks where one class significantly outnumbers the other. Class imbalance can lead to biased models that perform poorly on the minority class.

## Conclusion

The Credit Card Approval Prediction project successfully developed a machine learning model capable of predicting credit card approval decisions based on customer information. The project followed a structured methodology, involving data preprocessing, feature selection, model development, hyperparameter optimization, and comprehensive evaluation. The final model strikes a balance between predictive performance and interpretability, providing valuable insights for financial institutions in their credit card approval processes.

## Future Work and Recommendations

While the selected XG Boost model achieved satisfactory performance, there are opportunities for further improvement and extension of this project:

- Exploring additional feature engineering techniques or incorporating domain-specific knowledge to enhance model performance
- Investigating ensemble methods or stacking techniques to combine the strengths of multiple models
- Implementing online learning or updating mechanisms to adapt the model to evolving customer data and market conditions
- Integrating the model into a broader credit risk management system or decision support tool

## Contributions

The contributions of each team member to the project are as follows:

- Rithvik: Data acquisition, preliminary exploration, model evaluation.
- Alekhyaa: Team organization, project planning, hyperparameter optimization.
- Khyati: Data preprocessing, feature selection, initial model development.
- Nisarg: Data preprocessing, feature selection, final model evaluation.
- Jaswanth: Initial exploratory data analysis, model development, report preparation.

## References

- Raschka, Sebastian., Liu, Yuxi (Hayden)., Mirjalili, Vahid., Dzhulgakov, Dmytro. Machine Learning with PyTorch and Scikit-Learn: Develop Machine Learning and Deep Learning Models with Python. United Kingdom: Packt Publishing, 2022.
- Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., & Brunk, C. (1994). ReducingMisclassification Costs. InProceedings of the Eleventh International Conference onMachine LearningSan Francisco, CA. Morgan Kauffmann
- Koh, H. C., & Chan, K. L. G. (2002). Data mining and customer relationship marketing in the banking industry. Singapore Management Review
- S. B. Kotsiantis. Supervised Machine Learning: A Review of Classification Techniques. Informatica 31 (2007) 249-268 Web
- https://www.jair.org/index.php/jair/article/view/10302/24590
- https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9763647
- https://www.ijsce.org/wp-content/uploads/papers/v11i2/B35350111222.pdf
- https://www.ijraset.com/best-journal/credit-card-approval-prediction-using-classification-algorithms

**Presented By**

Rithvik HS
Khyati Desai
Nisarg Shah
Jaswanth Kumar Mannava
Alekhyaa Nelluri