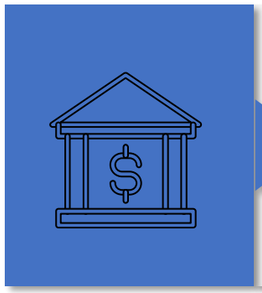
Credit Card Approval Prediction

Credit Scoring

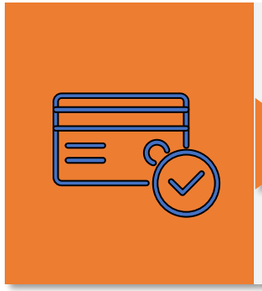
Assess creditworthiness using machine learning. Historical data predicts if an applicant is 'good' or 'bad'

Khyati Desai
4/25/2024
EM 589

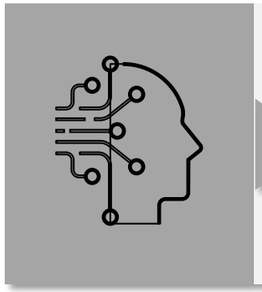
Introduction



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



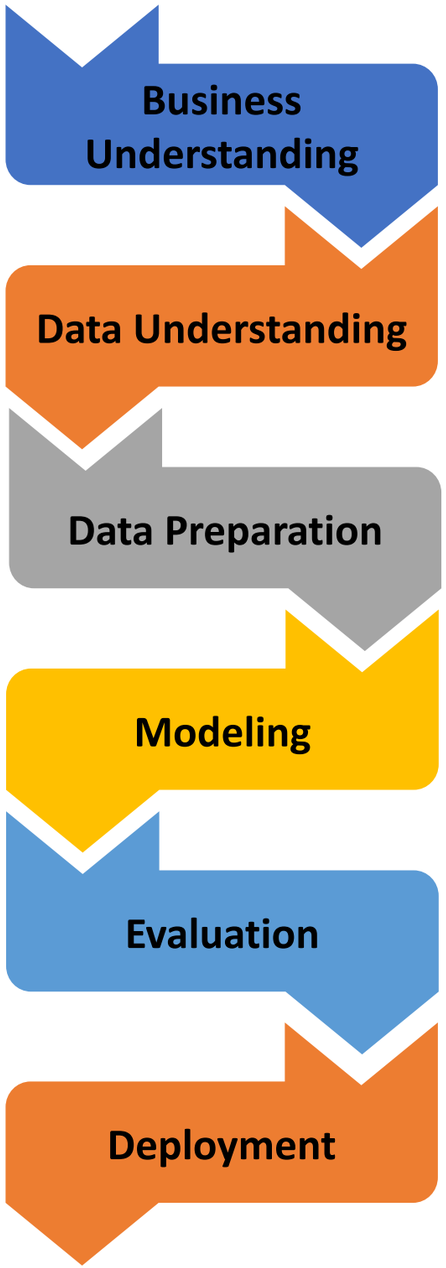
Automated credit approval prediction systems powered by machine learning algorithms offer a faster and more objective alternative



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

Khyati Desai
4/25/2024
EM 589

Project Roadmap



1. Exploring the potential of machine learning to expedite credit decisions and foster transparency
2. Unveiling hidden patterns in data crucial for precise credit card approval predictions
3. Creating and refining a clean dataset to improve the accuracy of predictive modeling
4. Customizing machine learning models to suit the unique characteristics of the dataset, optimizing their effectiveness in predicting credit card approvals
5. Conducted comparative analysis to identify and select the most effective model for credit evaluation
6. Integrating the most effective model seamlessly into banking systems to improve credit assessment processes and promote financial inclusion

Khyati Desai
4/25/2024
EM 589

Business Understanding

Leveraging data-driven predictive models to revolutionize the credit approval process, enhancing efficiency and accuracy

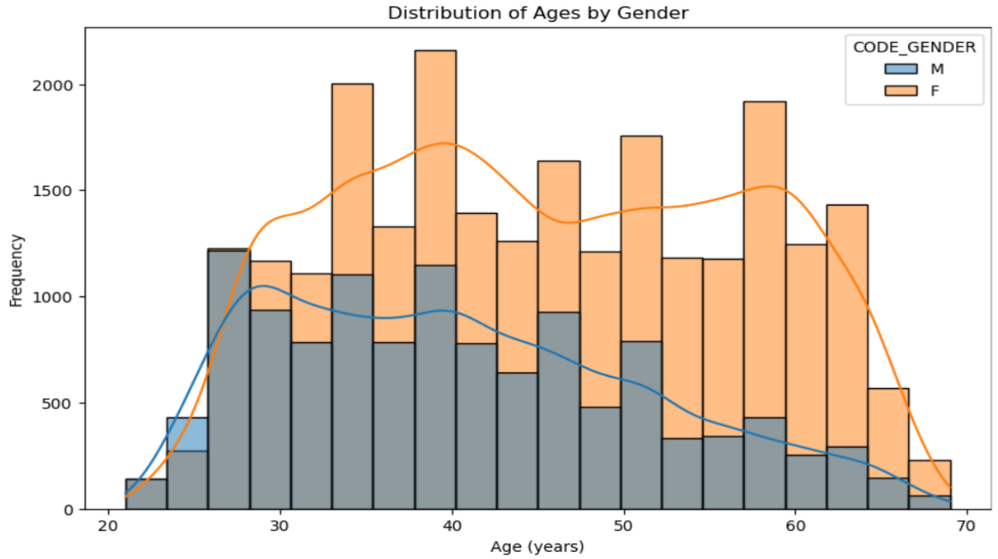
The banking sector struggles to accurately evaluate client creditworthiness, often relying on error-prone manual processes

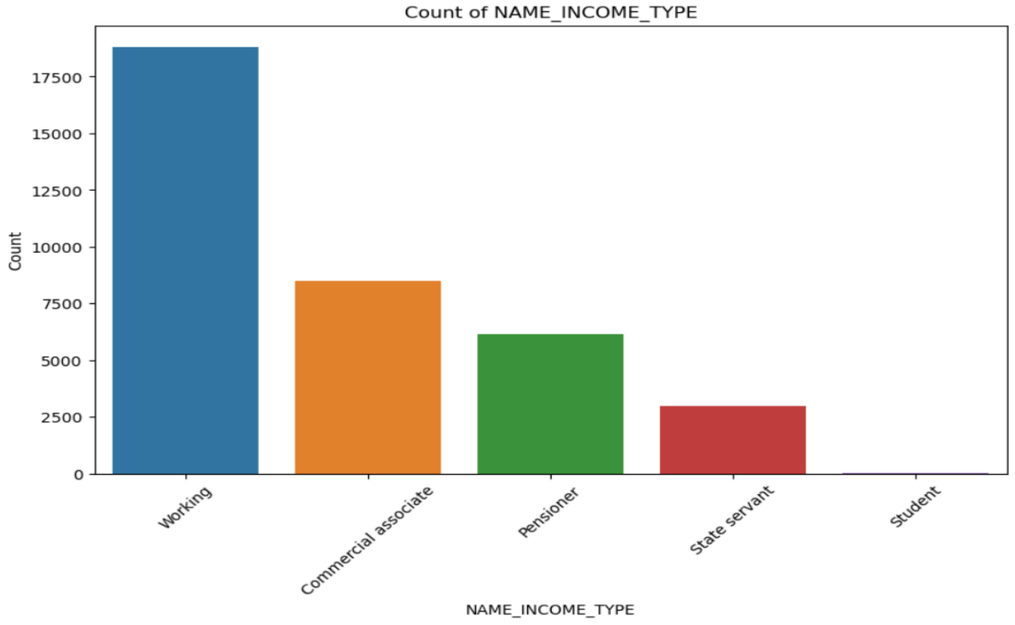
We develop accurate predictive models, providing insights to banks, fostering financial inclusion and empowerment

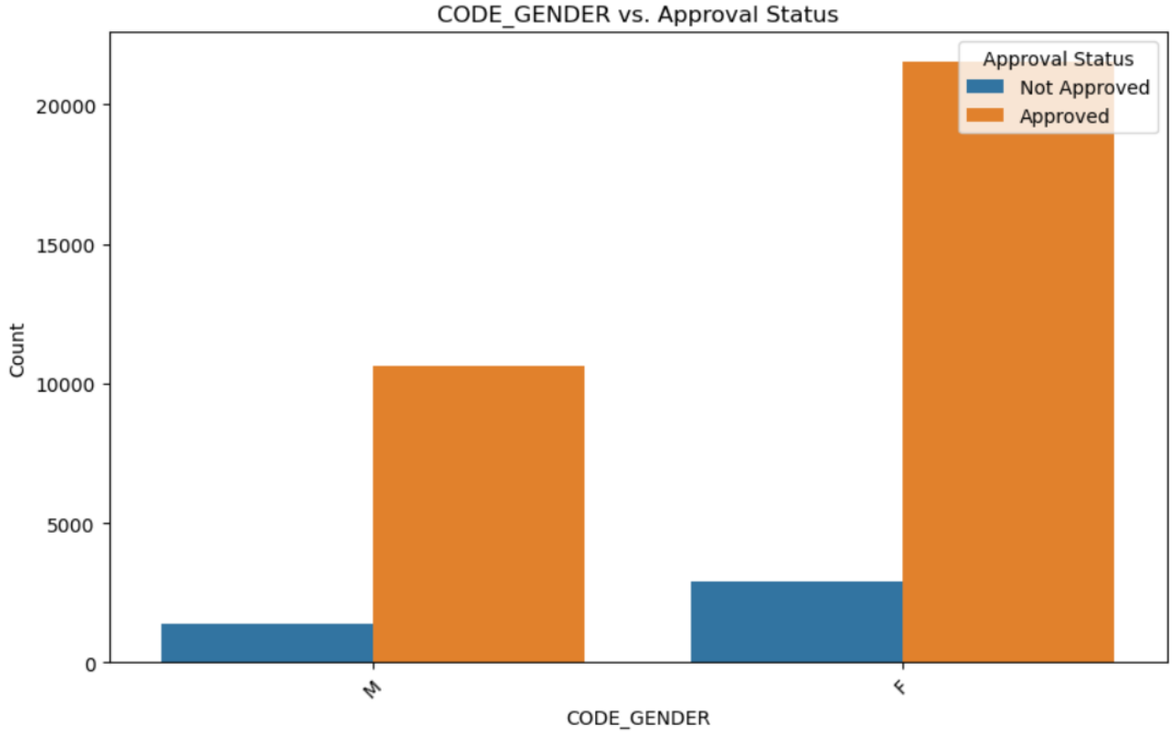
Using ML algorithms and historical data, we aim to predict credit card approvals accurately, optimizing decisions for financial institutions and empowering individuals financially

Khyati Desai
4/25/2024
EM 589

Exploratory Data Analysis







Skewed dataset – SMOTE

Khyati Desai
4/25/2024
EM 589

Data Preparation

Khyati Desai
4/25/2024
EM 589

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

Data Preparation

1. Credit Record

Columns:

- **ID:** A unique identifier for each record.
- **Months_Balance:** Indicates the number of months before the current month (e.g., 0 for the current month, -1 for the last month, and so on)
- **Status:** Represents the credit status, with codes ranging from 0 to 5, indicating different levels of overdue payments, and 'C' for paid off, and 'X' for no loan

	ID	MONTHS_BALANCE	STATUS
0	5001711	0	X
1	5001711	-1	0
2	5001711	-2	0
3	5001711	-3	0
4	5001712	0	C

Cleaning Procedure:

- Multiple entries exist for each ID due to different months of credit activity, ranging from the most recent month (0) to previous months (e.g., -1, -2, and so forth)
- Create a target column by consolidating the credit status values into a binary format (0 or 1) and selecting the maximum value among the categories
- Remove duplicate rows from the Credit Record dataset to streamline the data

Khyati Desai
4/25/2024
EM 589

2. Application Record

Columns:

- **Flag_Mobil, Flag_work_phone, Flag_email, Flag_phone:** These columns are dropped as they are not deemed relevant for analysis.
- **Name_Education_type:** Certain education types are grouped under 'No GED.'
- **Days_Birth:** Converted to Age_Years by dividing by 365 days.
- **Days_Employed:** Converted to Years_Employed (negative values are first inverted and then divided by 365 days), and values less than zero are discarded.
- **Name_Family_Status:** 'Civil Marriage' is replaced with 'Married.'
- **Cnt_Family_Members:** Only consider values less than or equal to 9.

Data Merging: Merge Application Record and Credit Record datasets based on the common ID column.

```
Index(['CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'CNT_CHILDREN',
       'AMT_INCOME_TOTAL', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE',
       'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE', 'DAYS_BIRTH',
       'DAYS_EMPLOYED', 'FLAG_MOBIL', 'FLAG_WORK_PHONE', 'FLAG_PHONE',
       'FLAG_EMAIL', 'OCCUPATION_TYPE', 'CNT_FAM_MEMBERS'],
      dtype='object')
```

Khyati Desai
4/25/2024
EM 589

Modeling

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

Khyati Desai
4/25/2024
EM 589

Feature Selection

Why RFE?

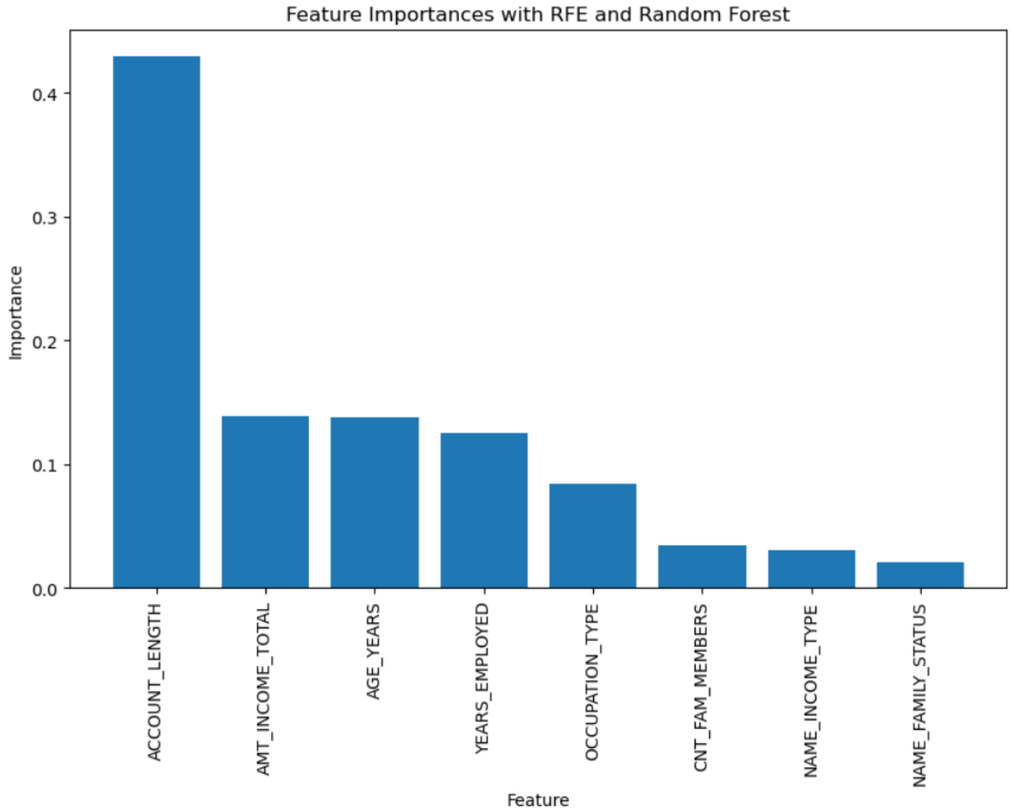
- Systematic evaluation of feature subsets, eliminating the least important features recursively
- Feature interactions: non-linear relationships
- Adaptability to various algorithms and high-dimensional datasets

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

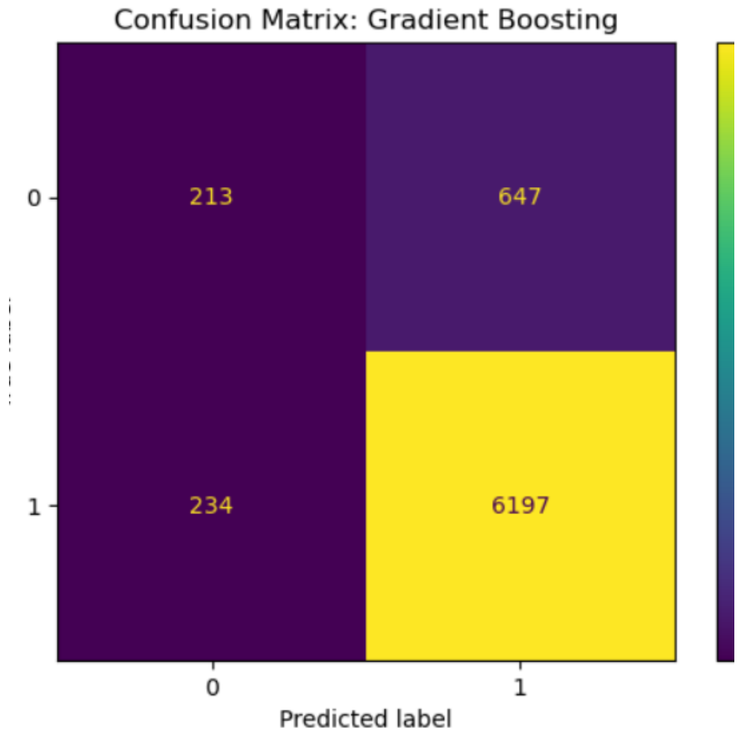
- RFE enhanced accuracy of RF model from 0.867 to 0.865 after feature selection
- Gradient Boosting Classifier: RFE improved accuracy from 0.87985 to 0.87917
- XGBClassifier: RFE improved accuracy from 0.88397 to 0.88136

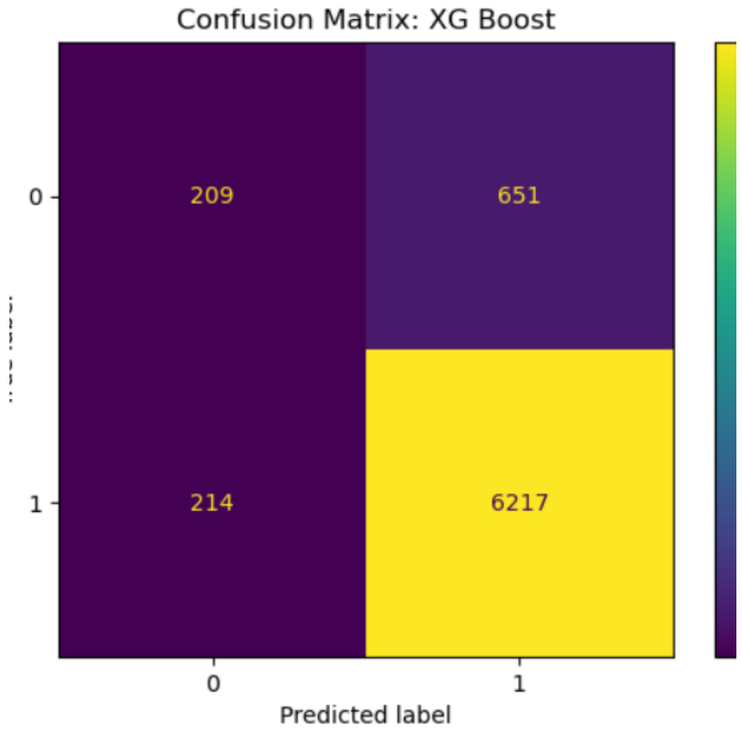


Khyati Desai
4/25/2024
EM 589

Evaluation

- Confusion matrices were generated to visualize the models' performance in classifying approved and not approved credit card applications

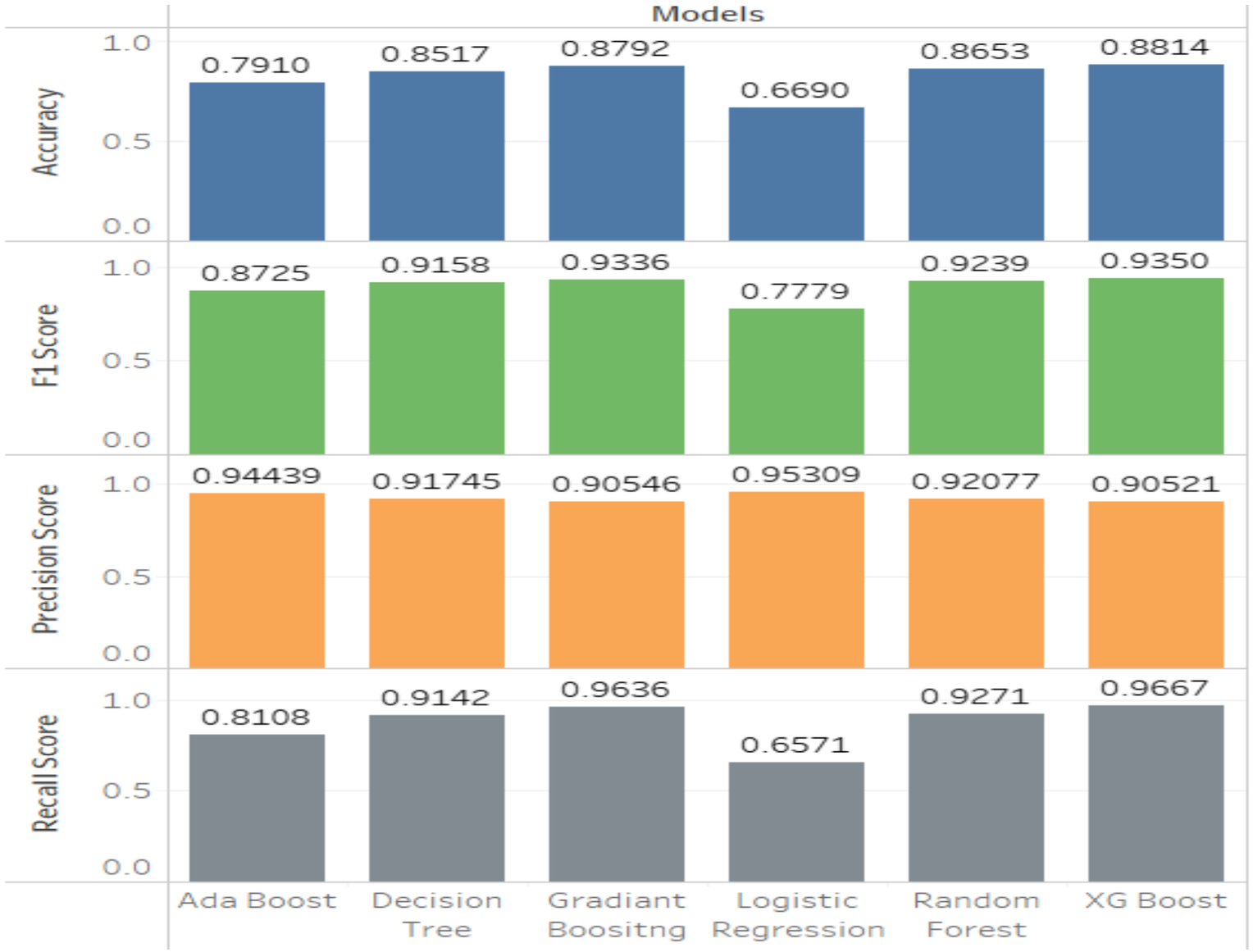




Khyati Desai
4/25/2024
EM 589

Evaluation

- The evaluation process involved assessing the models' accuracy scores, F1 scores, precision scores, and recall score



Khyati Desai
4/25/2024
EM 589

Deployment

1. Serialization

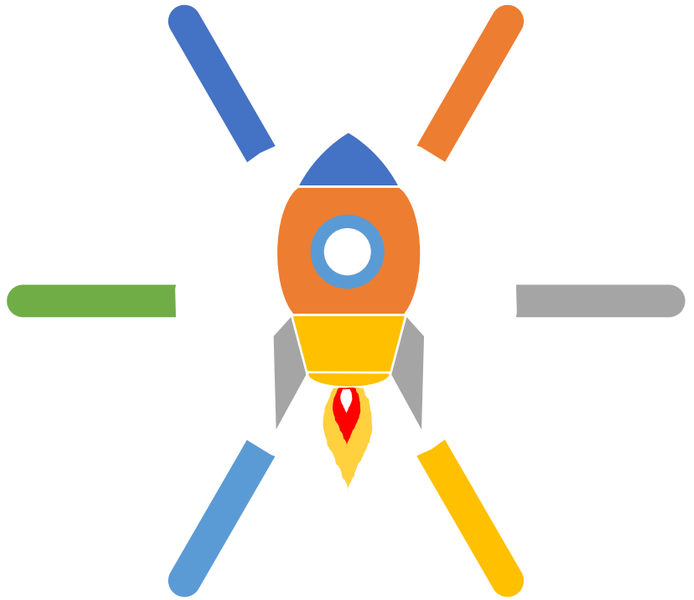
- Serialization of the selected model to save its parameters and structure for easy access

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

Khyati Desai
4/25/2024
EM 589

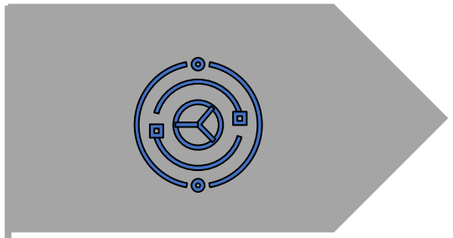
Conclusion

Comparable Accuracy

Both gradient boosting and XGBoost models achieved similar levels of accuracy, indicating that they are effective in capturing patterns in the data

Trade-off in Efficiency

While Gradient Boosting showed a very high accuracy, it came at the cost of significantly longer computational time compared to XGBoost



Linear vs. Non-linear Models

Logistic regression, displayed inferior performance compared to random forest, due to its linear relationship assumption between features and outputs.

Khyati Desai
4/25/2024
EM 589

Thank You!

-Contributors:

- Rithvik HS
- Khyati Desai
- Nisarg Shah
- Jaswanth Kumar Mannava
- Alekhyaa Nelluri