Artificial Intelligence and Machine Learning

Project Report

Semester-IV (Batch-2022)

CREDIT CARD APPROVAL PREDICTION

Supervised By:

Dr. Kirandeep Singh

Submitted By:

Aryan Walia

Ayushmaan

Ayush Thakur

Bahaar Sharma

Bhumi Ahlawat

Department of Computer Science and Engineering
Chitkara University Institute of Engineering & Technology,
Chitkara University, Punjab

1. INTRODUCTION

1.1 BACKGROUND

In recent years, with the proliferation of financial services and the ease of access to credit, the process of credit card approval has become increasingly critical for both financial institutions and potential clients. The decision-making process behind credit card approval involves a myriad of factors, including an applicant's credit history, income level, debt-to-income ratio, and other demographic information. Traditionally, this process has been subjective and prone to human biases, leading to inefficiencies and potential discrimination.

Evolution of Credit Card Approval Process

The credit card approval process has undergone significant evolution over the years, mirroring advancements in technology and changes in consumer behavior. Initially, credit decisions relied heavily on manual assessment, where bank officers would evaluate applicants based on limited information such as income and assets. However, this approach was often subjective and prone to errors, leading to inefficiencies and potential discrimination.

Emergence of Machine Learning in Finance

In recent decades, the emergence of machine learning (ML) techniques has revolutionized various industries, including finance. ML algorithms have proven adept at analyzing vast amounts of data to uncover complex patterns and relationships that might not be apparent to human analysts. In the realm of credit risk assessment, ML models offer the promise of more accurate predictions and better risk management.

Need for Objective Credit Scoring Models

Despite the advancements in credit scoring techniques, challenges persist in creating objective and unbiased models. Traditional credit scoring models often rely on predetermined rules and thresholds to classify applicants as 'good' or 'bad' clients. However, these rules may not capture the nuances of individual financial situations and can lead to arbitrary decisions. There is a growing need for ML-based models that can autonomously learn from data and adapt to changing credit landscapes.

1.2 OBJECTIVES

The primary objective of this project is to develop a machine learning model capable of predicting whether an applicant is likely to be a 'good' or 'bad' client based on the available data. Unlike traditional approaches where the definition of 'good' or 'bad' is predetermined, our model aims to autonomously identify patterns and characteristics associated with creditworthiness. By harnessing the power of machine learning algorithms, particularly the Random Forest classification technique, we seek to create a predictive model that can assist financial institutions in making more informed and objective decisions regarding credit card approvals.

Development of Predictive Model

The primary objective of this project is to develop a predictive model that can accurately assess the creditworthiness of applicants. Unlike traditional credit scoring models, which rely on predefined criteria, our model aims to learn from historical data to identify patterns associated with creditworthiness. By leveraging the power of ML algorithms, particularly Random Forest classification, we seek to create a more flexible and adaptive model capable of handling complex decision-making scenarios.

Mitigation of Human Bias

Another key objective is to mitigate the inherent biases present in traditional credit scoring processes. Human decision-makers may unconsciously introduce biases based on factors such as race, gender, or socioeconomic status, leading to unfair outcomes. By employing ML algorithms that operate on objective data, we aim to reduce the influence of subjective judgments and promote fairness and inclusivity in the credit approval process.

Enhancement of Risk Management

Furthermore, we aim to enhance risk management practices within financial institutions by providing more accurate predictions of credit risk. A robust predictive model can help lenders identify high-risk applicants and take proactive measures to mitigate potential losses. By improving the accuracy of credit risk assessments, our model can contribute to the overall stability and sustainability of the financial system.

1.3 SIGNIFICANCE

The significance of this project lies in its potential to revolutionize the credit card approval process, making it more efficient, transparent, and fair. By leveraging machine learning, we can mitigate the inherent biases present in human decision-making, leading to a more inclusive and equitable financial system. Furthermore, an accurate predictive model can help financial institutions minimize the risk of default and improve overall portfolio performance. Ultimately, the successful implementation of our credit card approval prediction model can have far-reaching implications for both lenders and borrowers, fostering trust and reliability in the credit industry.

Transformation of Credit Industry

The successful implementation of our credit card approval prediction model has the potential to transform the credit industry by streamlining the approval process and making it more data-driven. By automating credit decisions and reducing reliance on subjective judgments, financial institutions can improve operational efficiency and reduce processing times. This transformation can enhance the overall customer experience and foster trust and confidence in the lending process.

Inclusivity and Fairness

Moreover, our model aims to promote inclusivity and fairness in credit access by minimizing the impact of biased decision-making. By treating each applicant as an individual and evaluating their creditworthiness based on objective criteria, we can reduce disparities in access to credit and ensure equal opportunities for all. This commitment to fairness aligns with regulatory objectives and societal expectations regarding responsible lending practices.

Risk Mitigation and Portfolio Performance

From a risk management perspective, our predictive model offers tangible benefits for financial institutions by helping them identify and manage credit risk more effectively. By identifying high-risk applicants early in the approval process, lenders can implement risk mitigation strategies such as adjusting interest rates or imposing stricter lending criteria. This proactive approach can ultimately lead to lower default rates, improved portfolio performance, and enhanced shareholder value.

In the subsequent sections of this report, we will delve into the methodology employed to develop our predictive model, the implementation details, the results obtained, and a thorough discussion of the findings, limitations, and future directions

2. PROBLEM DEFINATION AND REQUIREMENTS

2.1 PROBLEM STATEMENT

Complexity of Credit Approval Decision

The process of credit card approval is inherently complex, involving the assessment of multiple factors to determine an applicant's creditworthiness. Traditional approaches to credit scoring rely on predefined criteria to classify applicants as 'good' or 'bad' clients based on thresholds set by financial institutions. However, these criteria may not capture the full spectrum of individual financial situations and can lead to inaccurate or biased decisions. Our objective is to develop a predictive model that can autonomously learn from historical data to make more nuanced and accurate credit approval decisions.

Challenge of Subjectivity and Bias

One of the key challenges in credit card approval is the presence of subjective judgments and biases in the decision-making process. Human decision-makers may inadvertently introduce biases based on factors such as race, gender, or socioeconomic status, leading to unfair outcomes. By leveraging machine learning techniques, we aim to mitigate these biases by basing credit decisions on objective data rather than subjective assessments.

Need for Adaptive and Data-Driven Approaches

In today's dynamic financial landscape, traditional credit scoring models may struggle to adapt to changing market conditions and consumer behaviors. Our approach seeks to address this challenge by developing a predictive model that can autonomously learn from data and adapt to evolving credit landscapes. By harnessing the power of machine learning algorithms, particularly Random Forest classification, we aim to create a flexible and adaptive model capable of making accurate credit approval predictions in real-time.

2.2 SOFTWARE REQUIREMENTS

Python for Machine Learning Development

Python is chosen as the primary programming language for machine learning development due to its versatility, extensive libraries, and active community support. With libraries such as scikit-learn, pandas, and numpy, Python provides a comprehensive ecosystem for data preprocessing, model development, and evaluation.

Utilization of Machine Learning Libraries

Scikit-learn, a widely used machine learning library in Python, offers a diverse range of algorithms and tools for classification, regression, clustering, and dimensionality reduction. We will leverage scikit-learn's implementation of the Random Forest algorithm for building our credit card approval prediction model.

Interactive Development Environment

Jupyter Notebook is selected as the preferred development environment for its interactive and exploratory nature, allowing for seamless integration of code, documentation, and visualizations. Jupyter Notebook provides an ideal platform for iterative model development, enabling us to experiment with different algorithms and parameter settings efficiently.

2.3 DATA SETS

Comprehensive Historical Credit Data

The primary dataset comprises historical credit card application data, including applicant demographics, financial information, credit history, and approval outcomes. This dataset serves as the foundation for training and evaluating the predictive model, providing insights into patterns and trends associated with creditworthiness.

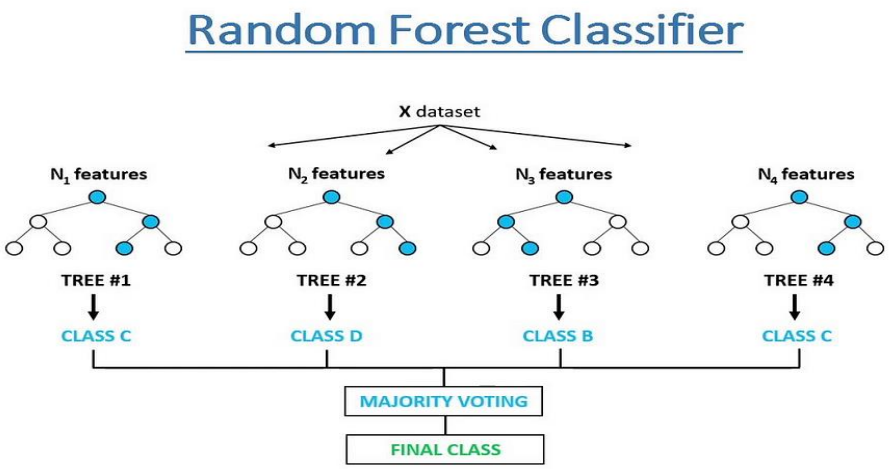
Supplementary Data for Enhanced Predictions

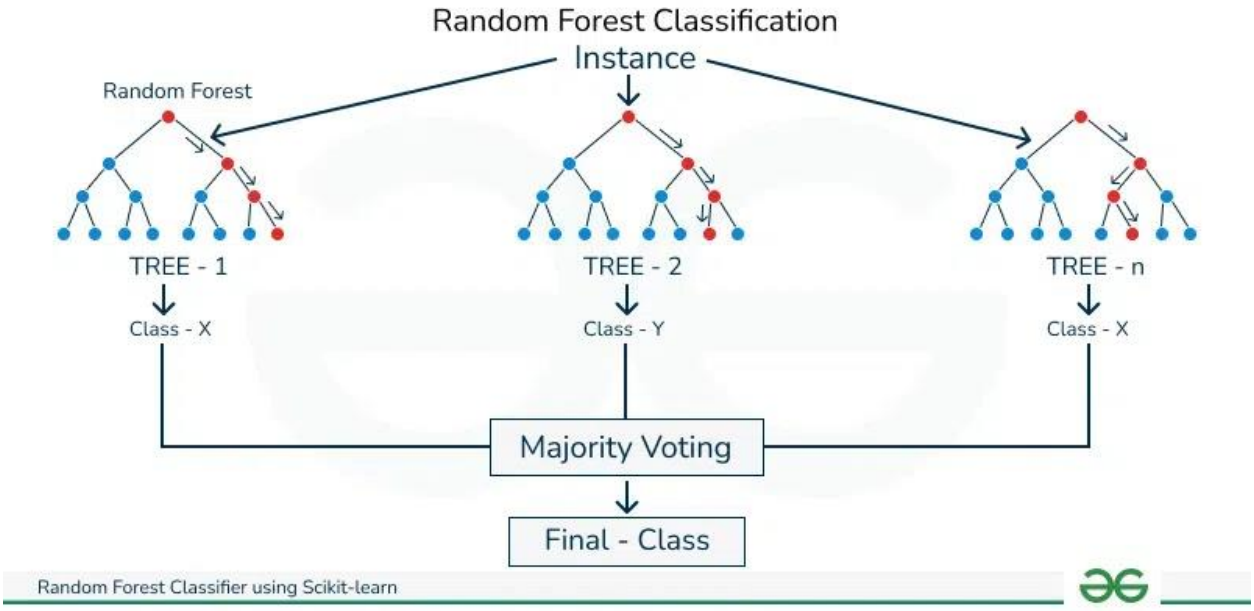
In addition to the primary dataset, supplementary data containing external factors that may influence creditworthiness can be incorporated to enrich the predictive model. Economic indicators, industry trends, and regulatory changes are examples of supplementary data sources that can enhance the model's predictive power and robustness.

In the subsequent sections, we will delve into the methodology employed to address the identified problem statement, including data preprocessing, feature engineering, model development, and evaluation.

3. PROPOSED DESIGN AND METHODOLOGY

3.1 SCHEMATIC DIAGRAM





3.2 ALGORITHMS USED

1. Random Forest Classifier (RFC):

- Purpose: The RFC model is employed for binary classification to predict whether a credit card applicant is a 'good' or 'bad' client.
- Description: Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of individual trees. It combines the predictions of multiple decision trees to improve accuracy and reduce overfitting.
- Implementation: Implemented using the RandomForestClassifier class from the scikit-learn library in Python.

2. Data Resampling Techniques:

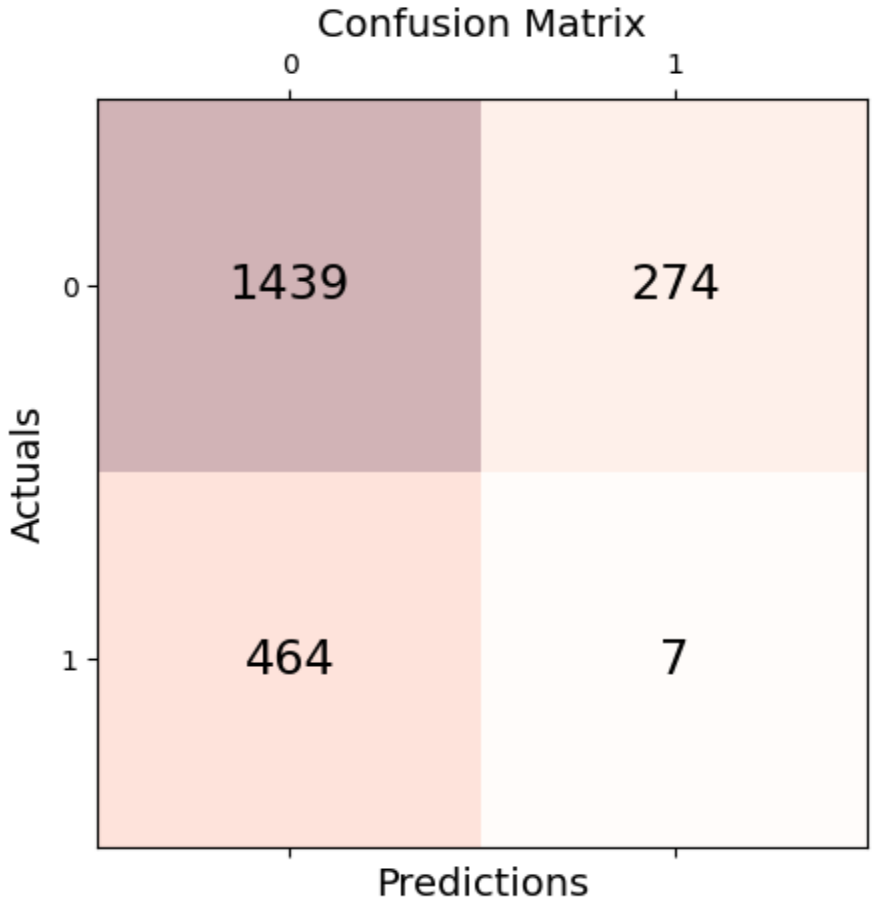
- Purpose: Used to address class imbalance in the dataset, where the number of 'good' clients significantly outweighs the number of 'bad' clients.
- Description: Resampling techniques such as Random Under Sampling (RUS) are applied to balance the class distribution by either reducing the majority class or increasing the minority class.
- Implementation: Utilized the RandomUnderSampler class from the imbalanced-learn library in Python to perform random under-sampling.

3. Feature Importance Analysis:

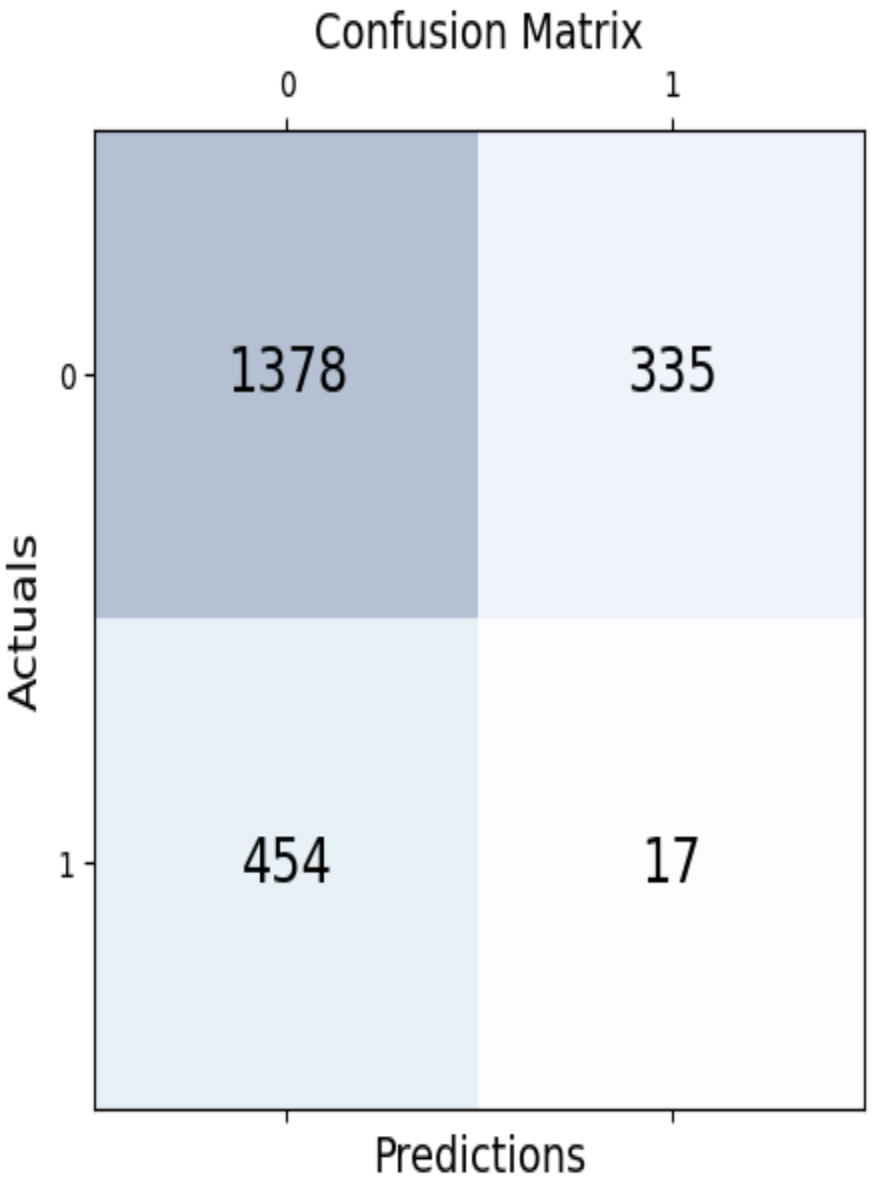
- Purpose: To identify the most influential features in the prediction of credit card approval.
- Description: Feature importance analysis evaluates the contribution of each feature in the model's decision-making process. It provides insights into which features have the most significant impact on the prediction outcome.
- Implementation: Calculated using the Mean Decrease in Impurity (MDI) method, which measures the decrease in node impurity averaged over all decision trees in the forest. Implemented using feature_importances_ attribute of the RandomForestClassifier in scikit-learn.

4.RESULTS

The model got things right about 87.68% of the time. When it said "yes," it was usually right, around 90.31% of the time. And when the model tried to catch the right answers, it did pretty good, about 96.12% of the time. However, it was a bit tricky for the model to say "no" correctly. It only got about 32.15% of those right. The overall result, F1 score, was 93.12%.



Accuracy: 66.21%
Precision: 2.49%
Recall: 1.49%
Specificity: 84.00%
F1: 1.86%



Accuracy: 63.87%

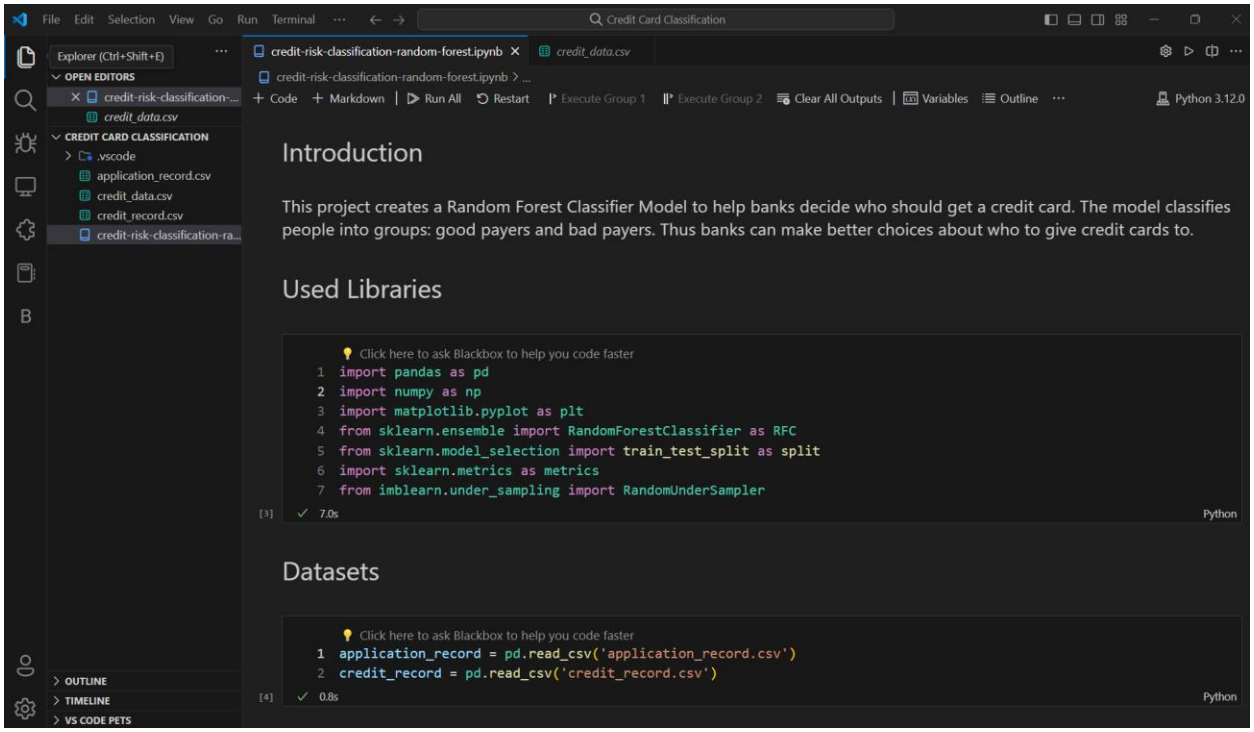
Precision: 4.83%

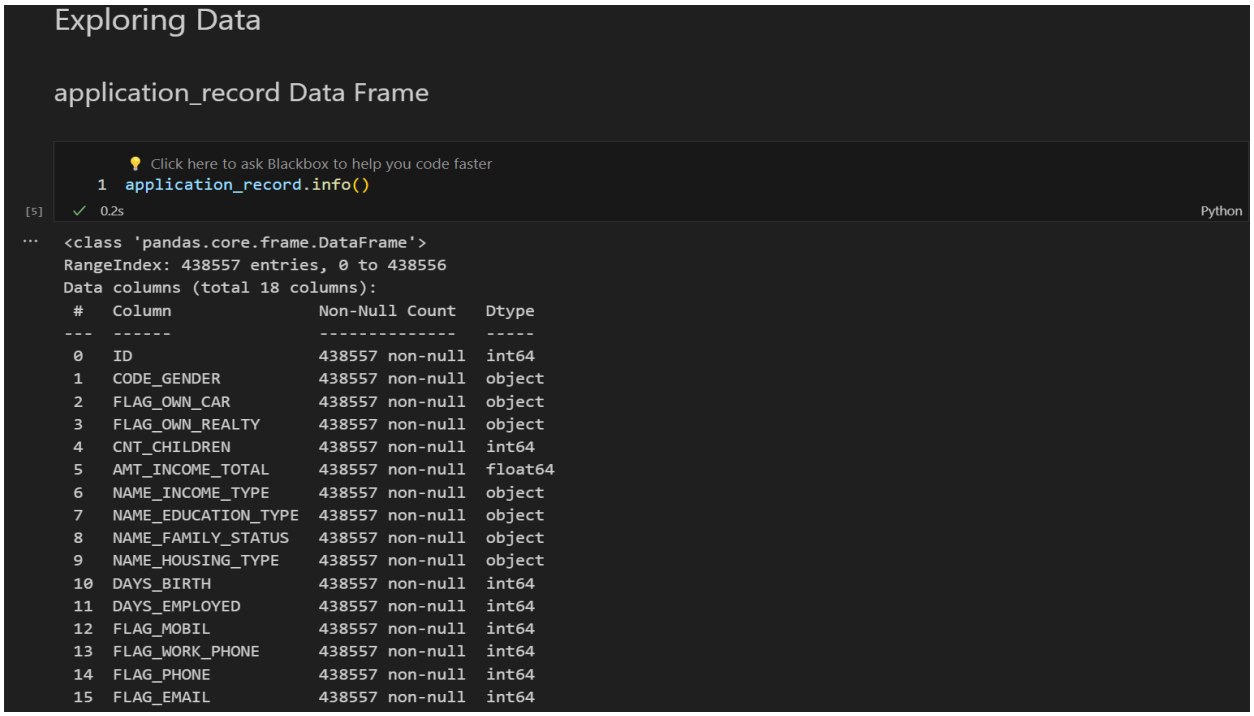
Recall: 3.61%

Specificity: 80.44%

F1: 4.13%

4.1 PROJECT SCREENSHOTS





The application_record dataset gathers 438,557 observations regarding the personal and financial information of the bank's customers. Each row represents a specific bank customer, having unique ID number, and contains 17 different information about them.

```
Click here to ask Blackbox to help you code faster
1 application_record.nunique()
```

[6] ✓ 0.2s Python

```
ID                   438510
CODE_GENDER               2
FLAG_OWN_CAR              2
FLAG_OWN_REALTY           2
CNT_CHILDREN             12
AMT_INCOME_TOTAL        866
NAME_INCOME_TYPE          5
NAME_EDUCATION_TYPE       5
NAME_FAMILY_STATUS        5
NAME_HOUSING_TYPE         6
DAYS_BIRTH            16379
DAYS_EMPLOYED          9406
FLAG_MOBIL                1
FLAG_WORK_PHONE           2
FLAG_PHONE                2
FLAG_EMAIL                2
OCCUPATION_TYPE          18
CNT_FAM_MEMBERS          13
dtype: int64
```

Treating Duplicates

```
Click here to ask Blackbox to help you code faster
1 # Check for duplicate values in ID column
2 duplicates_bool = application_record.duplicated(subset='ID', keep=False)
3 print('There are',sum(duplicates_bool),'duplicates in ID column.')
```

[7] ✓ 0.0s Python

```
There are 94 duplicates in ID column.
```

```
Click here to ask Blackbox to help you code faster
1 duplicates = application_record[duplicates_bool].sort_values('ID')
2 duplicates.head(10)
```

[8] ✓ 0.0s Python

	ID	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	NAME_INCOME_TYPE
426818	7022197	M	Y	Y	3	135000.0	Working
425023	7022197	F	N	Y	0	450000.0	Commercial associate
431545	7022327	F	N	Y	0	135000.0	Commercial associate
431911	7022327	M	Y	Y	0	256500.0	Commercial associate
425486	7023108	M	Y	Y	1	67500.0	Working
426488	7023108	F	N	N	0	135000.0	Working

This data pertains to 45,985 bank customers and encompasses up to 60 months of retroactive credit card billing history for each customer.

Dependent Variable

The definition of 'good' or 'bad' cleint not given. A model that predicts whether a customer's profile would lead to a credit card approval, the payment history of each customer was categorized into three clusters: "Not Enough Data" (cluster -1), "Bad Payer" (cluster 0), and "Good Payer" (cluster 1).

Analysing Credit Decision

"Not Enough Data" category (cluster -1) were those who hadn't utilized their credit card at all, indicating a lack of transaction history. 4,536 customers (excluded)

"Bad Payers" (cluster 1) had delayed payments of over 30 days at least once in their payment history. 5,350 customers.

"Good Player" (status "C") or payments up to 29 days late (status "0") were categorized as low risk and considered suitable profiles for credit card approval (cluster 0). 36,099 customers

```
💡 Click here to ask Blackbox to help you code faster
#new dataset "credit_decision" with two columns:
#the unique customer IDs with credit card history and classification decisions (-1, 0, or 1).
credit_decision = pd.DataFrame()
credit_decision['ID'] = credit_record['ID'].unique()
```

[13] ✓ 0.0s Python

When compared, customers with the same ID have very different information and even opposite genders.

credit_record Data Frame

```
💡 Click here to ask Blackbox to help you code faster
credit_record.info()
```

[11] ✓ 0.0s Python

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1048575 entries, 0 to 1048574
Data columns (total 3 columns):
 #   Column          Non-Null Count    Dtype
---  ------          --------------    -----
 0   ID              1048575 non-null  int64
 1   MONTHS_BALANCE  1048575 non-null  int64
 2   STATUS          1048575 non-null  object
dtypes: int64(2), object(1)
memory usage: 24.0+ MB
```

The dataset named "credit_record" contains information about customer payment history across different months, totaling 1,048,575 entries.

```
💡 Click here to ask Blackbox to help you code faster
credit_record.nunique()
```

[12] ✓ 0.0s Python

Checking Imbalance Data

```
Click here to ask Blackbox to help you code faster
1 credit_decision['Decision'].value_counts(normalize=True)
```

[48] ✓ 0.0s Python

```
Decision
0    0.870926
1    0.129074
Name: proportion, dtype: float64
```

```
Click here to ask Blackbox to help you code faster
1  fig, ax = plt.subplots()
2
3  client = ['Good Clients (Decision = 0)','Bad Clients (Decision = 1)']
4  proportions = credit_decision['Decision'].value_counts(normalize=True)
5  bar_colors = ['tab:green','tab:red']
6
7  ax.bar(client, proportions, color=bar_colors)
8
9  ax.set_ylabel('Percentage of Dataset')
10 ax.set_title('Data Imbalance')
11
12 ax.text(1-0.1, proportions[1]/2, '{:.2%}'.format(proportions[1]), size=10)
13 ax.text(0-0.1, proportions[0]/2, '{:.2%}'.format(proportions[0]), size=10)
14
15 fig.show()
```

Independent Variables

The features of this dataset: 1. continuous numeric variables,2. Numeric categorical and 3.Non-numerical categorical variables. The non- numerical categorical variables were required to be transformed into numerical ones.

Categorical Features

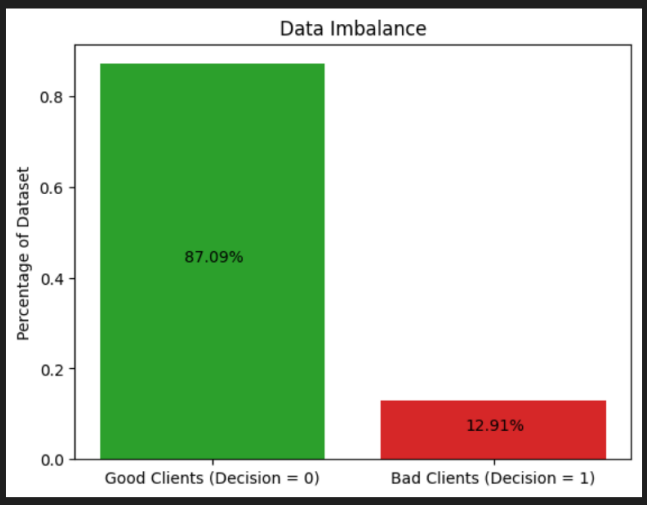
Factorizing categorical variables allows the Random Forest Classifier to assemble decision trees based on numerical rules (larger or smaller). Such a model can deal well with numeric categorical variables without needing them to be transformed into dummy variables.

```
Click here to ask Blackbox to help you code faster
1  # Separates the categorical variables
2  categorical_columns = ['FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE', 'NAME_FAMILY_S
3  categorical_data = credit_data[categorical_columns]
4
5  # Factorizes categorical variables transforming them into numerical data
6  numerical_categorical_data = categorical_data.apply(lambda x: pd.factorize(x)[0])
7  numerical_categorical_data = pd.DataFrame(numerical_categorical_data)
8
9  numerical_categorical_data.head()
```

[46] ✓ 0.0s Python

	FLAG_OWN_CAR	FLAG_OWN_REALTY	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS	NAME_HOUSING_T
0	0	0	0	0	0	0

```
fig.show()
```




```
Click here to ask Blackbox to help you code faster
credit_data.to_csv('credit_data.csv')
```

[50] ✓ 0.1s Python

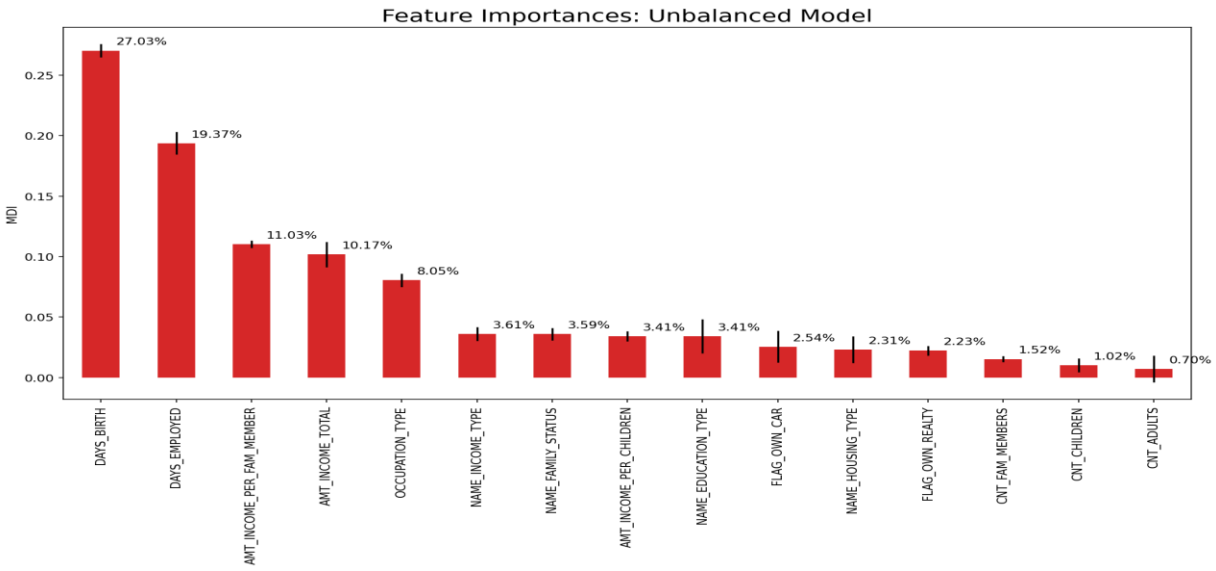
```
fig.show()
```

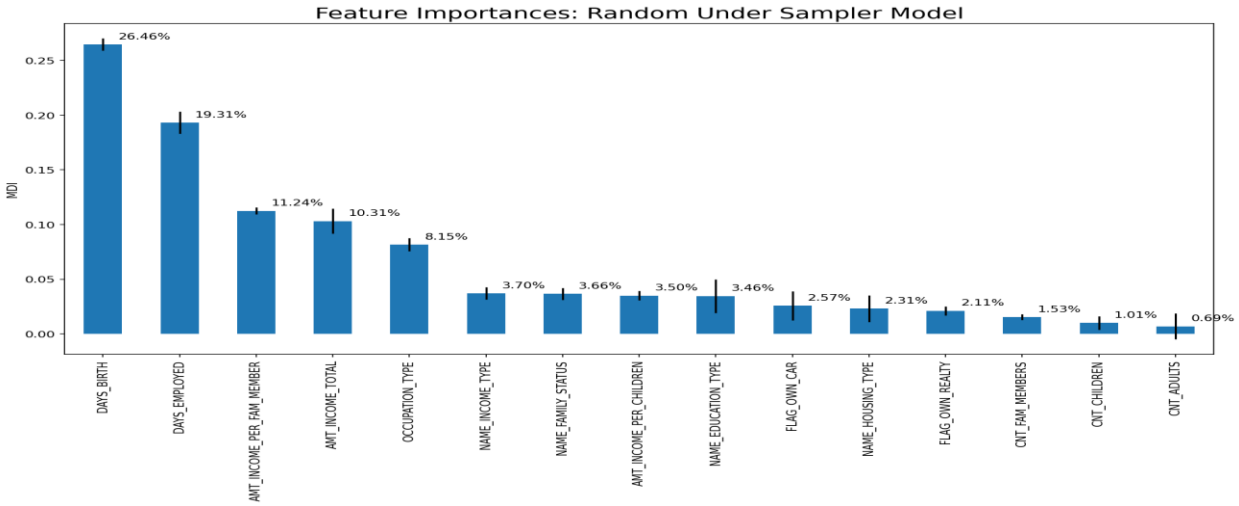

```
Click here to ask Blackbox to help you code faster
unbal_rf = RFC( n_estimators = 1000, max_features = 8, random_state=0)
unbal_rf.fit(unbal_X_train, unbal_y_train)
```

[54] ✓ 1m 9.5s Python

4. FEATURE IMPORTANCE

According to Mean Decrease in Impurity (MDI) analysis, the DAYS_BIRTH feature has the most importance (25.12%), followed by DAYS_EMPLOYED (18.44%), AMT_INCOME_TOTAL (15.99%) and OCUPPATION_TYPE (8.19%). All other features have an MDI of less than 5%. The FLAG_MOBIL feature has special irrelevance in the model's prediction capacity since every customer has a registered cell phone, making its importance equal to 0%.





5.CONCLUSION

The model did a good job overall. It was accurate most of the time, getting things right around 87.68% of the cases. It was particularly good at saying "yes" to the right people, with about 90.31% of its "yes" answers being correct. It also caught many of the real "yes" cases, about 96.12%. However, it struggled a bit in saying "no" correctly, with only 32.15% of its "no" answers being right. The overall score, called F1, was 93.12%, showing it did well in balancing everything.

One thing to remember is that the dataset was not balanced since 87% of clients were classified as good payers while only 13% were considered bad profile. Also, the dataset only had people who actually got the card, not those who were denied at first for other reasons, like unpaid debts. This might affect the model's performance in real situations. In the future, it would be a good idea to improve the model's ability to say "no" more accurately and consider these imbalances while making the model better.