CREDIT CARD APPROVAL CAPSTONE PROJECT REPORT

PROJECT INTRODUCTION

A bank's credit card department is one of the top adopters of data science. A top focus for the bank has always been acquiring new credit card customers. Giving out credit cards without doing proper research or evaluating applicants' creditworthiness is quite risky. The credit card department has been using a data-driven system for credit assessment called Credit Scoring for many years, and the model is known as an application scorecard. A credit card application's cutoff value is determined using the application scorecard, which also aids in estimating the applicant's level of risk. This decision is made based on strategic priority at a given time.

Customers must fill out a form, either physically or online, to apply for a credit card. The application data is used to evaluate the applicant's creditworthiness. The decision is made using the application data in addition to the Credit Bureau Score, such as the FICO Score in the US or the CIBIL Score in India, and other internal information on the applicants. Additionally, the banks are rapidly taking a lot of outside data into account to enhance the caliber of credit judgements.

PROJECT OBJECTIVE

The main objective of this assignment is to minimize the risk and maximize the profit of the bank. Bank has to make a decision based on the applicant's profile to minimize the loss from the bank's perspective. Bank considers the applicants over their nature of work, income range and family orientation details to take any decision to approve or reject a credit card application. The customer Credit card data contains many features and a classification approach to identify the credit worthiness of an applicant.

In this project we are utilizing the exploratory data analysis (EDA) as a data exploration technique to acquire knowledge, discover new relations, apply new methodologies and unravel patterns in data. It is important to apply the necessary rationale behind each step to address the main objective of the study.

DATASET DESCRIPTION

Ind ID	Client ID
Gender	Gender information
Car owner	Having car or not

Property owner	Having property or not
Children	Count of children
Annual income	Annual income
Type income	Income type
Education	Education level
Marital status	Marital status
Housing type	Living style
Birthday count	Use backward count from current day (0), -1 means yesterday.
Employed days	Start date of employment. Use backward count from the current day(0). positive value means the individual is currently unemployed
Mobile phone	Any mobile phone
Work phone	Any work phone
Phone	Any phone number
EMAIL_ID	Any email ID
Type Occupation	Occupation
Family Members	Family size
Credit card label ID	The joining key between application data and credit status data, same is Ind_id
Credit card label Label	0 is application approved and 1 is application rejected.

SECTION 1:

1.IMPORTANCE OF THE PROPOSAL(CREDITCARD APPROVALPREDICTION) IN TODAY'S WORLD

Predicting the creditworthiness of clients is crucial for banks in today's world for several reasons:

- **Risk Management:** It helps banks make safer lending decisions by identifying reliable borrowers, reducing the risk of unpaid debits.

- **Efficiency:** Speeds up the credit approval process, making it faster and more convenient for customers.
- **Customer Targeting:** Ensures that banks offer credit cards to individuals who are more likely to use them responsibly, improving customer satisfaction and bank profits.
- **Competitive Edge:** Gives banks an advantage by making their credit decisions smarter and quicker.
- **Data-Driven Decisions:** Uses data to make informed credit decisions, which can uncover hidden insights and trends.
- **Customization:** Allows banks to tailor their credit decisions to match their risk tolerance and business goals.
- **Financial Inclusion:** Extends credit to individuals who may not have an extensive credit history but are trustworthy borrowers.
- **Regulatory Compliance:** Helps banks follow rules and regulations by providing a transparent and consistent approach to credit assessment.

2. IMPACT ON BANKING SECTOR:

Creating a strong credit assessment system makes a big difference in the banking sector:

- **Profitability:** By reducing loan defaults and attracting more creditworthy customers, these model scan significantly boost a bank's profitability.
- **Targeted Marketing**: Data-driven models enable banks to tailor credit offers to individuals who are more likely to qualify. This targeted marketing approach increases the likelihood of conversions and revenue growth.
- **Customer Experience**: Faster credit decisions translate to a better customer experience. Clients appreciate streamlined processes and are more likely to remain loyal to the bank.
- **Risk Mitigation:** A robust credit assessment model helps banks identify creditworthy borrowers accurately, reducing the risk of loan defaults. This, in turn, safeguards the bank's financial stability.
- **Portfolio Quality**: Effective credit assessment models improve the overall quality of the bank's loan portfolio by ensuring that loans are granted to borrowers who are more likely to repay them.

3. GAP IN KNOWLEDGE AND FUTURE BENEFITS IN BANKING IN INDIA:

One potential gap in current banking practices in India is the reliance on traditional credit assessment methods that may not fully leverage the power of data and advanced analytics. Many banks still use rule-based credit scoring systems that primarily consider historical financial data. This gap in knowledge pertains to the underutilization of modern machine learning techniques and non-traditional data sources in credit assessment.

How the Proposed Method Can Be Helpful in the Future for Any Bank in India:

Implementing advanced machine learning-based credit assessment models can bridge this knowledge gap and provide substantial benefits to banks in India:

Enhanced Accuracy: Machine learning models can analyze a broader range of data, including transaction history, social media activity, and more. This leads to more accurate creditworthiness assessments, reducing the risk of default.

Financial Inclusion: By considering a wider array of factors, these models can extend credit to individuals with limited credit histories or unconventional financial backgrounds, promoting financial inclusion among underserved populations.

Customization: Banks can customize these models to align with their specific risk tolerance and lending strategies, making credit decisions more adaptable to diverse customer segments.

Operational Efficiency: Automation through machine learning streamlines the credit approval process, reducing manual effort and making lending operations more efficient.

Regulatory Compliance: Implementing transparent machine learning models can aid banks in complying with evolving regulations, ensuring fairness, and reducing legal risks.

Improved Portfolio Quality: By accurately identifying creditworthy borrowers, banks can improve the overall quality of their loan portfolios, leading to better financial health.

So, the proposal to predict credit card approval is highly relevant in today's banking sector, offering numerous benefits such as improved risk management, enhanced efficiency, and targeted customer engagement. This innovation bridges the gap in traditional credit assessment methods by leveraging advanced machine learning and diverse data sources. Its impact includes increased profitability, better customer experiences, and broader financial inclusion, making it a valuable addition to the future of banking in India.

SECTION 2:

1.INITIAL HYPOTHESIS FOR DATA ANALYSIS (DA) TRACK:

1.How does annual income impact credit card approval rates

Hypothesis: Higher annual income is positively correlated with higher credit card approval rates. Individuals with higher income levels are more likely to

be approved for credit cards as they have the financial means to repay debts.

2.Is there a relationship between education level and credit card approval?

Hypothesis: Education level may or may not influence credit card approval rates. Applicants with higher education levels may have better financial literacy, leading to higher approval rates.

3.Does marital status affect credit card approval?

Hypothesis: Marital status might influence credit card approval. Married individuals may have access to shared financial resources, potentially impacting their creditworthiness.

4.How does the number of family members relate to credit card approval?

Hypothesis: The number of family members could affect credit card approval. Larger families may have higher expenses and, thus, different spending behaviours that impact creditworthiness.

5.Does the type of income influence credit card approval rates?

Hypothesis: The type of income source may affect credit card approval. Stable income sources like salaries may result in higher approval rates compared to irregular income sources.

6.Does property ownership relate to credit card approval?

Hypothesis: The property ownership may affect credit card approval. Property ownership will help to analyse the stability a person who will be applying for credit.

7.Is there a relationship between employment status and credit card approval?

Hypothesis: Employment status is something similar to education level which may or may not be affective. Employed person have the high possibility for getting the approval of credit card.

2.INITIAL HYPOTHESIS FOR MACHINE LEARNING (ML) TRACK:

1.Can we build an effective machine learning model to predict credit card approval status based on applicant information?

Hypothesis: It is possible to build a predictive machine learning model that uses applicant information, such as annual income, education, marital status, and more, to accurately classify credit card approval status. The model's performance can be evaluated using relevant evaluation metrics.

2.What is the optimal machine learning algorithm for credit card approval prediction?

Hypothesis: Different machine learning algorithms, such as Logistic Regression, Random Forest, Support Vector Machine (SVM), and Decision Tree, may perform differently in predicting credit card approval. We hypothesize that one of these algorithms will outperform the others in terms of accuracy, precision, and recall.

3.Can we identify the most important features that impact credit card approval decisions?

Hypothesis: Feature importance analysis can help identify the most influential factors that impact credit card approval decisions. We hypothesize that annual income, age, and type of income may be among the most important features.

4.Is the predictive model generalizable to new data, and does it outperform random chance?

Hypothesis: The machine learning model can be generalized to new, unseen data, and its performance will be significantly better than random chance. This can be justified by comparing model performance metrics, such as accuracy, precision, recall, and Fl-score, to random guessing.

5.How do different hyperparameter settings affect the machine learning model's performance?

Hypothesis: Hyperparameter tuning can significantly impact the machine learning model's performance. We hypothesize that optimizing hyperparameters using techniques like Grid Search will result in improved model performance compared to default settings.

These initial hypotheses provide a starting point for data analysis and machine learning experiments, aiming to uncover insights and build an effective predictive model for credit card approval.

DATA ANALYSIS APPROACH:

1.Approach to Prove or Disprove Hypotheses: In order to prove or disprove the initial hypotheses, we will follow these steps:

- **Exploratory Data Analysis (EDA):** We will begin with EDA to gain insights into the dataset. We will visualize and analyse the distributions of variables, correlations, and potential outliers. EDA will help us identify initial patterns and relationships in the data.
- **Feature Engineering:** Based on EDA findings, we may create new features or transform existing ones to capture relevant information. For example, we might create income brackets or encode categorical variables.
- **Visualizations:** We will use various data visualizations, including bar plots, histograms, and scatter plots, to illustrate important patterns and relationships identified during EDA.

2.Feature Engineering Techniques: Relevant feature engineering techniques for this project may include:

- **Encoding**: Converting categorical variables (e.g. education, marital status) into numerical format using techniques like one-hot encoding.
- **Standardisation:** Scaling numerical variables to ensure they have similar scales, which can be important for some machine learning algorithms.

3 Justification of Data Analysis Approach: Our approach is justified for the following reasons:

- **Understanding the Data:** EDA is essential to understand the datasets characteristics, distributions, and potential outliers. It helps us identify the data's limitations and peculiarities.
- **Feature Engineering**: Feature engineering enhances the dataset's quality by creating meaningful variables and preparing the data for modelling.
- **Visualization:** Visualizations are powerful tools to communicate findings and support decision-making. They help in presenting patterns and relationships in a clear and interpretable manner.

4.Identifying Important Patterns Using EDA: During the EDA process, we aim to identify important patterns and relationships in the data, such as:

- **Missing Values:** Identifying Missing values that affect the analysis and treating them all.
- **Outliers:** Identifying outliers that may impact the analysis and may need special handling.
- **Feature Importance:** Assessing the importance of different features in predicting credit card approval. This helps in selecting relevant features for modelling.

Overall, our data analysis approach aims to provide a comprehensive understanding of the dataset, validate hypotheses, and prepare the data for machine learning modelling to predict credit card approval effectively.

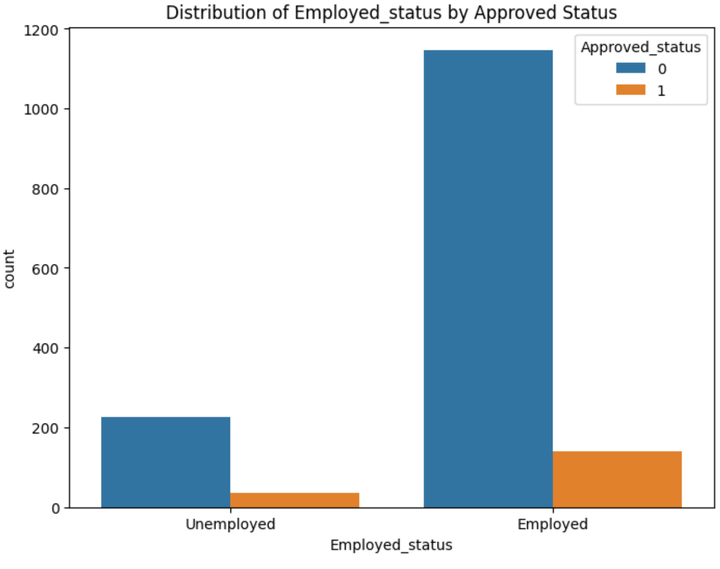
DATA INSIGHTS

1.The total number of applicants who received approval for a credit card in the provided dataset.

Within the dataset, it is observed that roughly 89% of the applicants successfully obtained approval for the credit card, whereas the remaining 11% were not granted approval for the credit card, indicating a discernible disparity in approval rates.

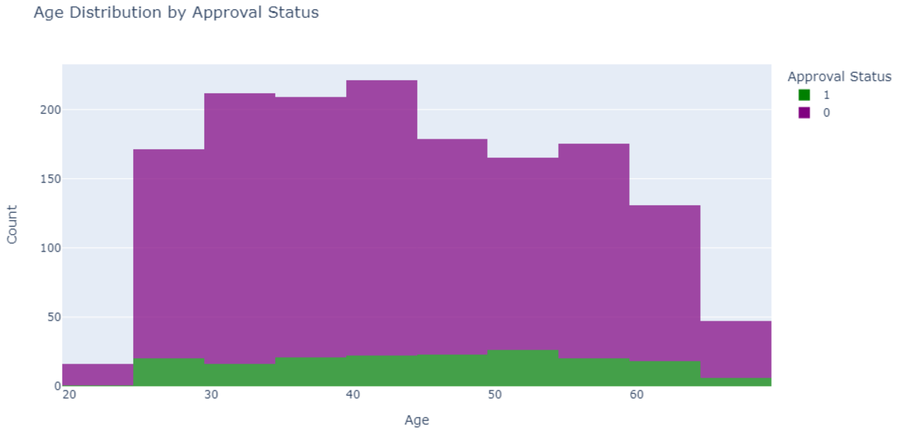
2.Exploring the correlation between employment status and approval status.

The analysis strongly confirms that people with jobs have the highest chance of getting approved for credit cards, just as we initially thought. Also, it's worth mentioning that some unemployed applicants who got approved might have other sources of income, like pensions or government jobs, which could explain why they were approved despite not having a regular job.



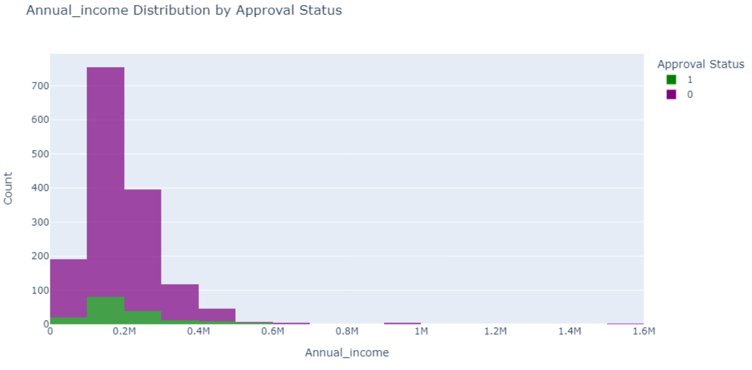
3.Analysing the age distribution concerning approval status.

In our analysis, we've found that a significant number of credit card applicants fall within the Age range of 25 to 60. Our visualizations clearly illustrate how credit card approval rates vary across different age groups, highlighting distinct approval patterns among applicants in this age range.



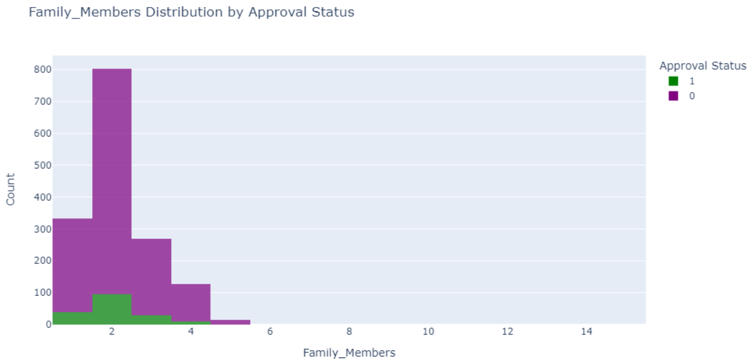
4.Annual Income Distribution by approval status

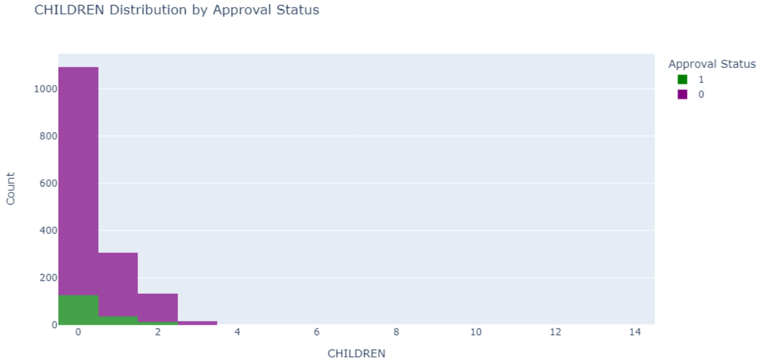
Based on our analysis, it's evident that individuals with annual incomes between 100,000 and 199,000 (1 Lakh) are more likely to secure credit card approvals. However, it's important to note that the dataset lacks currency information for earnings.



5.Analysing the relationship between family members, the presence of childern and their connection with approval status

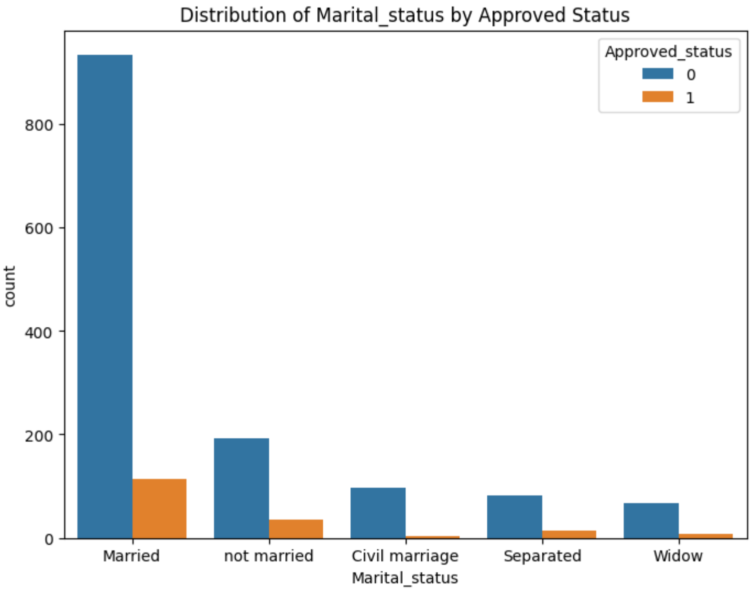
The data shows that a significant number of approved credit card applicants come from families with two members. This often indicates married couples without children. As per our initial hypothesis, it's reasonable to think that the number of family members can affect credit card approval. Larger families might have higher expenses, which can lead to different spending habits that influence their creditworthiness. This observation highlights the importance of considering family size as a significant factor in the credit card approval process.





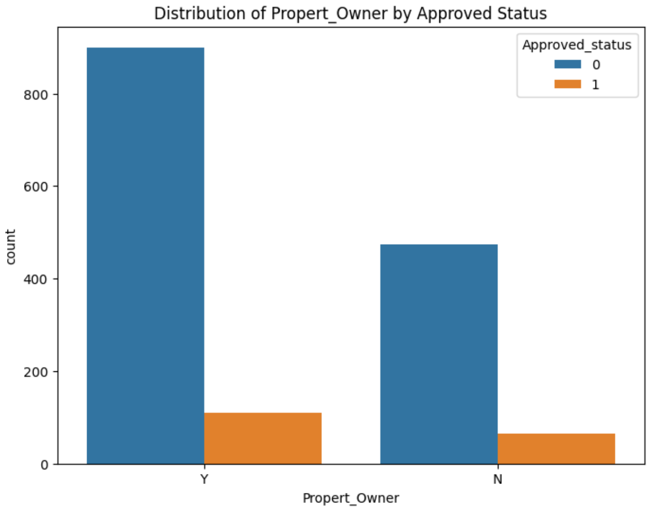
6.Analysing the connection between marital status and approval status.

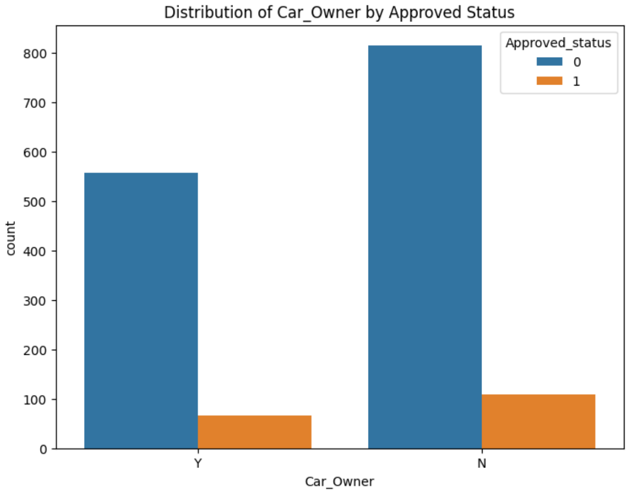
Our analysis has confirmed the initial hypothesis: a family size of 2 members frequently represents a married couple without children. Additionally, the data supports the observation that a larger percentage of married individuals have been granted credit card approvals. This aligns well with our initial assumption.



7.Analysing the relationship between the property and car ownership in relation to approval status

Based on our analysis, it's clear that owning a car or vehicle doesn't have a strong influence on credit card approval. Most applicants in our data don't own a car. However, property ownership stands out as a more significant factor. Those who own property, like a house or apartment, are more likely to get approved for a credit card. This suggests that owning property has a bigger impact on credit card approval than owning a vehicle.



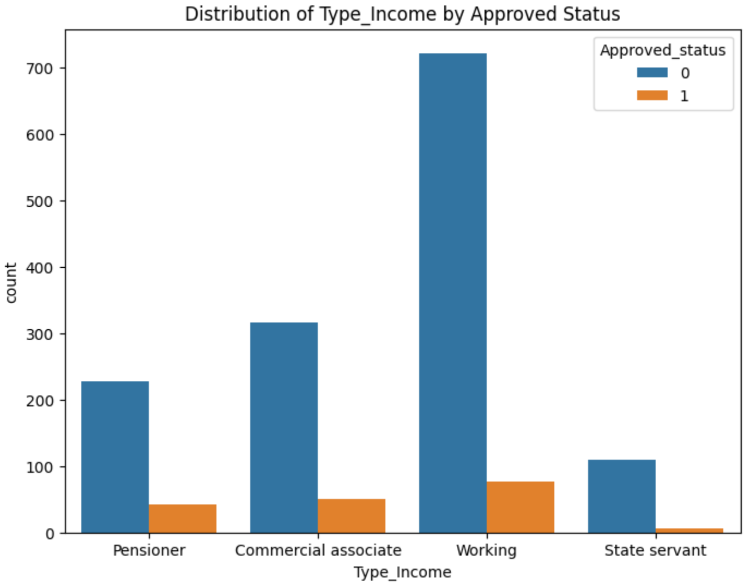


8.Analysing the connection between income type and approval status.

Based on our analysis, there's a noticeable association between income type and Status employment status within the dataset. Specifically:

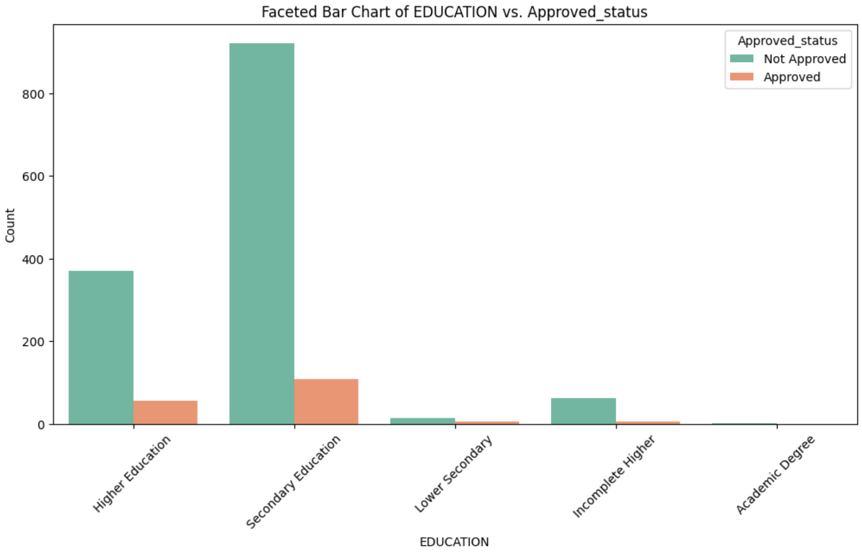
- **Applicants classified as "working," "commercial associate,” and "state servant” typically fall under the category of employed individuals.**
- **On the other hand, applicants labelled as "pensioner" are often categorized as unemployed.**

This observation strongly suggests a connection between the type of income and employment status among credit card applicants in our dataset



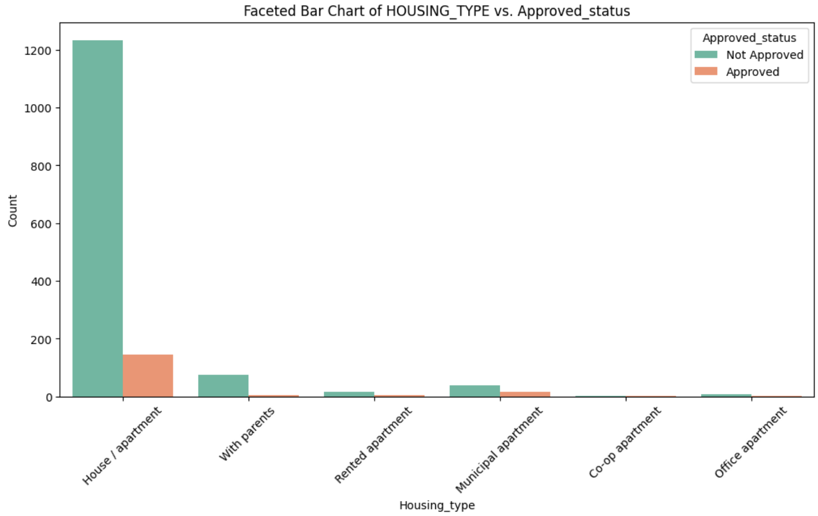
9.Analysing the correlation between education and approval status.

The data suggests that an applicant's education level may not have a strong impact on credit card approval. Most approved applicants have secondary education, and those with a degree have the lowest approval rates. However, it's worth noting that having passed secondary education appears favourable for approval.



10.Analysing the connection between housing type and approval status.

Many applicants live in 'house/apartment' type housing, and they are more likely to have their credit card applications approved compared to those with different housing arrangements.



Machine Learning Approach

1.Approach for Machine Learning Predictions:

a. Data Pre-processing:

- **Handle missing values: Ensure that there are no missing values in the dataset and apply mean and median techniques to fill missing values.**
- **Encode categorical variables: Convert categorical variables into numerical format using techniques like one-hot encoding.**

b. Data Splitting:

- **Split the dataset into training and testing sets. A common split is 80% for training and 20% for testing.**

c. Model Selection:

- **Choose machine learning algorithms suitable for classification tasks. Those models are Logistic Regression, Support Vector Machines, Decision Trees, Random Forests & XG Boost.**

d. Model Training:

- **Train the selected models on the training data.**

e. Model Evaluation:

- **Evaluate model performance on the testing data using appropriate metrics such as accuracy, precision, recall, and F1-score.**

f. Model Selection:

- **After evaluating the all the above-mentioned model, we'll select the most performing model for further process.**

2.Improving Model Accuracy through Essential Measures:

To improve the accuracy of model, I had taken steps to tune hyperparameters using

GridSearchCV for different algorithms.

a. Logistic Regression:

Best parameters: {'C": 10, 'penalty': 'l1', 'solver": 'liblinear'}

b. Support Vector Classifier:

Best parameters: {'C": 10, 'gamma’: 1, 'kernel": 'rbf'}

c. Decision Tree Classifier:

Best parameters: {'criterion'": 'gini', 'max_depth': 15}

d. Random Forest Classifier:

Best parameters: {'bootstrap': False, 'max_depth': 50, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 300}

e. XG Boost Classifier:

Best parameters: {‘colsample_bytree’:0.5,’learning_rate’: 0.1,’ max depth’: 20,’n_estimators’:

100,’subsample’1.0}

3.Assessing the Best Model Selection by Comparing Four Models:

To determine the best model, we should look at performance metrics and consider the goals of our credit card approval application. Let's analyze each model:

Models	Accuracy score
1.support vector classifier	**91.25%**
2.Desicion tree classifier	**93.26%**
3.Random forest classifier	**95.99%**
4.XG boost classifier	**96.72%**

Based on the performance metrics of the models (Support Vector Machine, Decision Tree, Random Forest, and XG Boost) and considering the goal of maximizing the accuracy of credit card approval prediction, the XG Boost Model emerges as the most suitable choice:

Accuracy Score

The XG Boost model achieved the highest accuracy score among the models, with a value of 96.72%. This indicates that XG Boost has learned well from the training data and exhibits excellent generalization performance, making it the top-performing model among those considered.

Precision Score

For predicting approved cases (0), the XG Boost model achieved a precision score of 0.97, signifying that when it predicted an approval, it was highly reliable. For predicting non-approved cases (1), it achieved a precision score of 0.97, indicating a high level of accuracy.

Recall Score

The XG Boost model demonstrated excellent recall for approved cases (0) with a score of 0.97, implying that it effectively captured nearly all actual credit card approvals. While its recall for non-approved cases (1) was slightly lower at 0.97.

F1 Score

The F1 Score, which combines precision and recall, was high for predicting approved cases (0), with a value of 0.97, indicating a good balance between precision and recall. For non-approved cases (1), the F1 Score was also 0.97, reflecting a trade-off between precision and recall, as expected.