**CREDIT CARD APPROVAL PREDICTION**

**Capstone 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 cut-off 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 calibre of credit judgements.

**OBJECTIVES:**

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: Features Name: *Credit\_Card.csv, Credit\_card\_label.csv***

**Ind\_ID:** Client ID

**Gender:** Gender information

**Car\_owner:** Having car or not

**Propert\_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 current day (0). Positive value means, 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.

**Why credit card approval Prediction is Necessary?**

1. **Importance of the Proposal in Today's World:**

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

1. **Risk Mitigation:** Banks face financial risks when granting credit cards to individuals. If the wrong individuals are approved, it can lead to defaults and financial losses. Accurate prediction of creditworthiness helps banks minimize these risks.
2. **Profitability:** Banks aim to generate profits from credit card operations. By identifying clients who are more likely to repay their debts on time, banks can maximize their profitability by offering credit cards to the right customers.
3. **Regulatory Compliance:** Regulatory bodies often require banks to adhere to certain standards and maintain a certain level of credit quality in their portfolios. Accurate credit assessment ensures compliance with these regulations.
4. **Customer Experience:** Approving credit cards for clients who are likely to manage them responsibly enhances the overall customer experience. Clients who can manage their credit well are more likely to remain loyal customers.
5. **Impact on the Banking Sector:**

Implementing an effective credit assessment model has a significant impact on the banking sector:

1. **Innovation:** Advancements in credit assessment techniques, such as machine learning and data analytics, have the potential to revolutionize the banking sector by providing more accurate predictions and insights.
2. **Reduced Risk:** Accurate credit assessment reduces the risk of loan defaults, which can have a cascading effect on a bank's financial health. It ensures that banks lend to clients who are more likely to repay their debts.
3. **Enhanced Profitability:** By targeting the right clients, banks can optimize their credit card portfolios. This optimization leads to higher profitability and improved financial performance.
4. **Competitive Advantage:** Banks with robust credit assessment models gain a competitive advantage. They can attract creditworthy clients and expand their customer base.
5. **Compliance:** Accurate credit assessment ensures that banks meet regulatory requirements. Non-compliance can result in fines and reputational damage.
6. **Future Relevance for Banks in India:**

The proposed method for credit assessment can be highly relevant for banks in India and other regions in the following ways:

1. **Risk Management**: As lending practices evolve, the need for effective risk management becomes more critical. Banks in India can benefit from advanced credit assessment methods to navigate the evolving credit landscape.
2. **Financial Inclusion:** India has a diverse population, and many individuals lack traditional credit histories. Innovative credit assessment methods can help banks extend credit to individuals who were previously underserved.
3. **Digital Transformation:** The Indian banking sector is undergoing digital transformation. Advanced credit assessment methods align with this transformation by leveraging data analytics and automation.
4. **Regulatory Compliance:** Regulatory requirements in India may evolve to address changing financial dynamics. A robust credit assessment model can help banks meet these requirements effectively.

In summary, the proposed method for credit assessment has the potential to address gaps in risk management, financial inclusion, and regulatory compliance in the Indian banking sector while leveraging digital innovations for enhanced performance and customer satisfaction.

**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.

1. 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.

1. 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.

1. 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.

1. 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.

1. 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.

1. 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.

**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.

1. 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.

1. 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.

1. 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 F1-score, to random guessing.

1. 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.

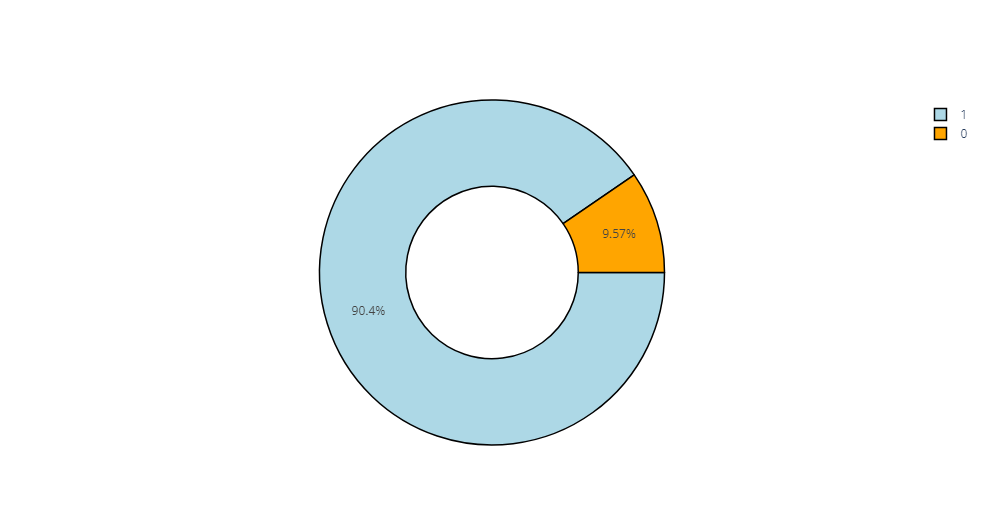
**Data Analysis Approach:**

1. **Approach to Prove or Disprove Hypotheses:** In order to prove or disprove the initial hypotheses, we will follow these steps:
2. ***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.
3. ***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.
4. ***Visualizations:*** We will use various data visualizations, including bar plots, histograms, and scatter plots, to illustrate important patterns and relationships identified during EDA.
5. **Identifying Important Patterns Using EDA**: During the EDA process, we aim to identify important patterns and relationships in the data, such as:
6. ***Missing Values:*** Identifying Missing values that affect the analysis and treating them all.
7. ***Outliers:*** Identifying outliers that may impact the analysis and may need special handling.
8. ***Feature Importance:*** Assessing the importance of different features in predicting credit card approval. This helps in selecting relevant features for modelling.
9. **Feature Engineering Techniques:** Relevant feature engineering techniques for this project may include:
10. ***Encoding:*** Converting categorical variables (e.g., education, marital status) into numerical format using techniques like one-hot encoding.
11. ***Standardisation:*** Scaling numerical variables to ensure they have similar scales, which can be important for some machine learning algorithms.
12. **Justification of Data Analysis Approach:** Our approach is justified for the following reasons:
13. ***Understanding the Data:*** EDA is essential to understand the dataset's characteristics, distributions, and potential outliers. It helps us identify the data's limitations and peculiarities.
14. ***Feature Engineering:*** Feature engineering enhances the dataset's quality by creating meaningful variables and preparing the data for modelling.
15. ***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.

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.

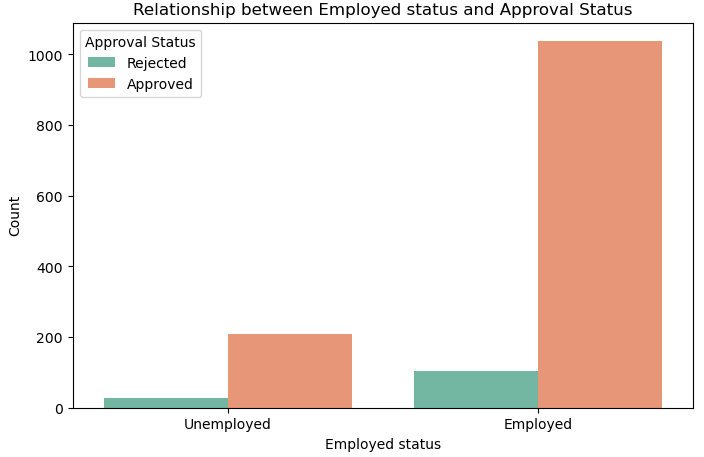
**Data Insights**

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

****

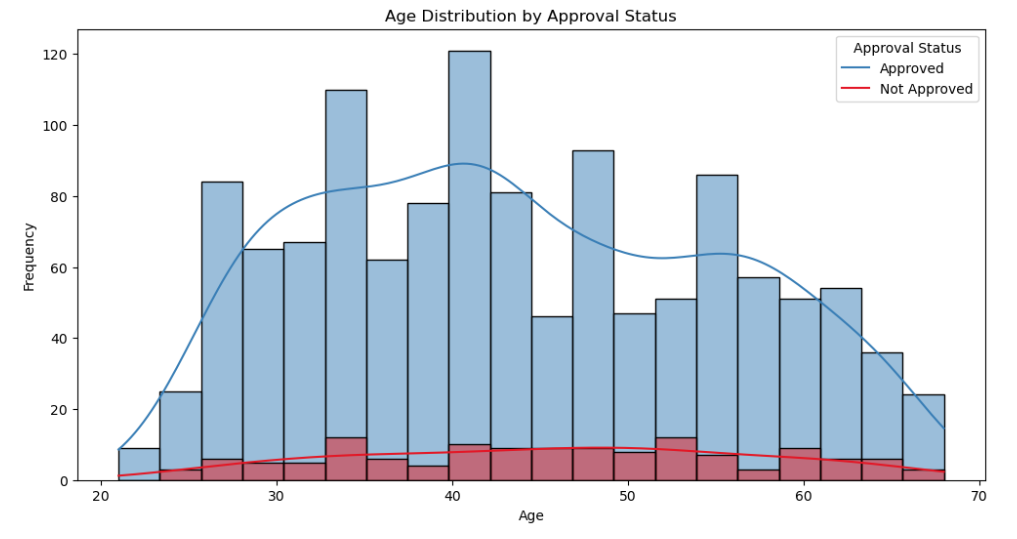
Within the dataset, it is observed that roughly 89% of the applicants successfully obtained approval for the credit card, whereas the remaining 11% were not granted approval for the credit card, indicating a discernible disparity in approval rates.

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



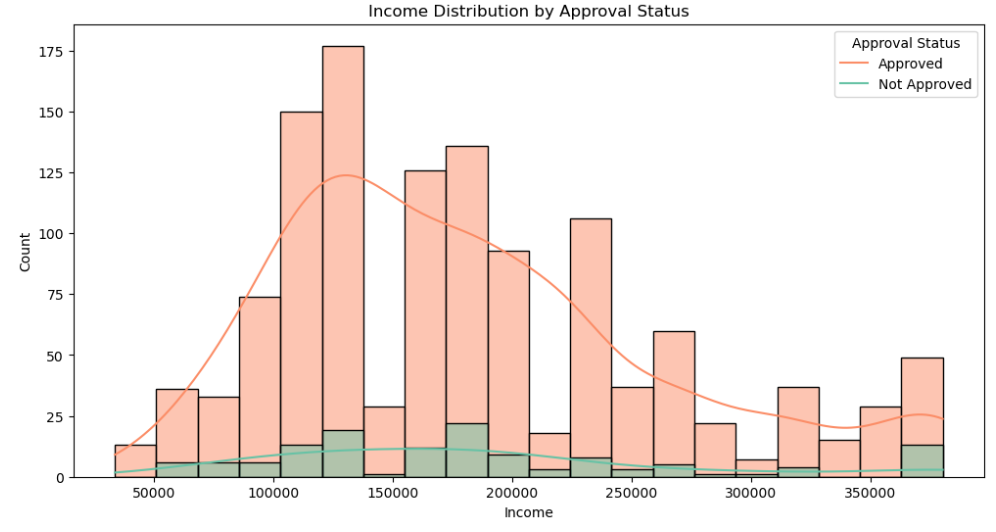
The analysis strongly supports the notion that employed applicants exhibit the highest rate of credit card approval, aligning with the initial hypothesis. Furthermore, it's important to note that certain unemployed applicants who were approved for credit cards may have alternative income sources, such as being pensioners, state servants, or commercial associates. These additional income sources potentially explain their credit card approval despite their official unemployment status.

1. Analysing the age distribution concerning approval status.



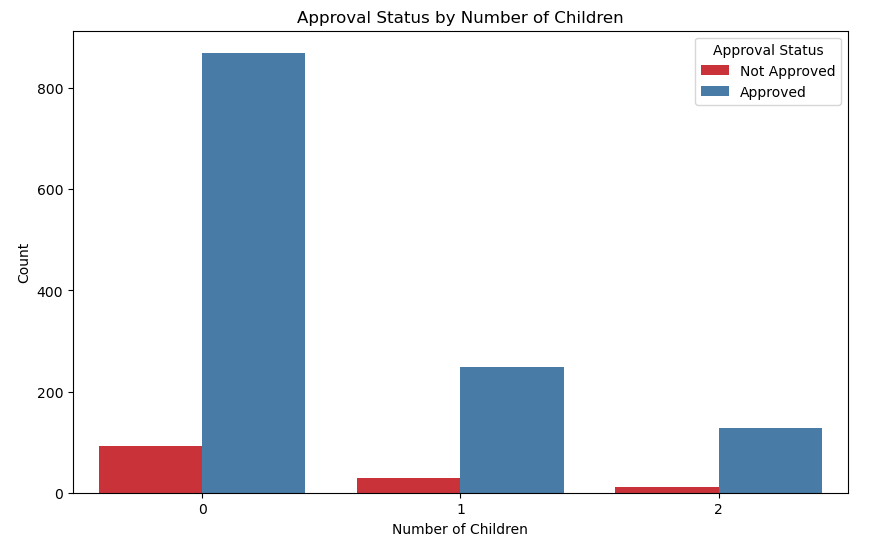
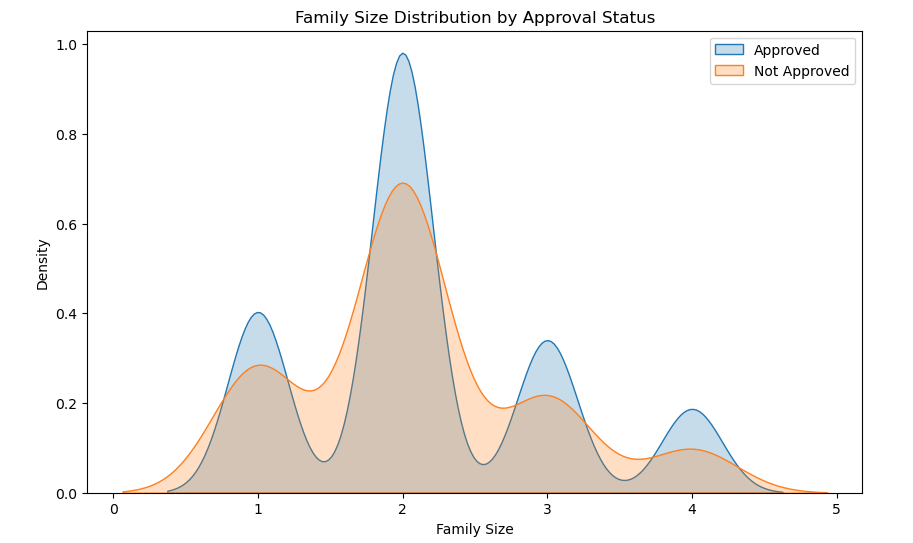
In our analysis, it becomes evident that a substantial proportion of individuals applying for credit cards tend to cluster within the age bracket spanning from 30 to 60 years old. Moreover, the visual representations provided above offer a comprehensive overview of the approval status distribution among various age groups, thus shedding light on the credit card approval patterns with respect to different age demographics.

1. Annual Income Distribution by approval status



The dataset analysis suggests that applicants with earnings exceeding 1 Lakh are more likely to obtain credit card approvals. However, it's important to acknowledge the absence of information regarding the currency used for earnings. Nonetheless, the data strongly indicates that a higher income is a contributing factor in the approval of credit card applications, aligning with our hypothesis.

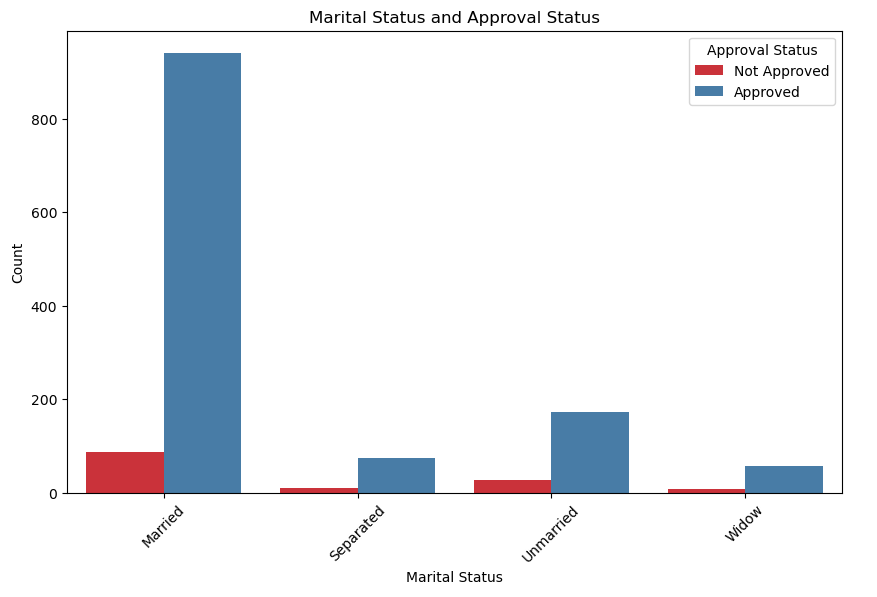
1. Analysing the relationship between family size, the presence of children, and their connection with approval status.



The data clearly indicates that a substantial percentage of approved credit card applicants fall within the family size category of 2. This trend is suggestive of married individuals without children, a common classification for couples.

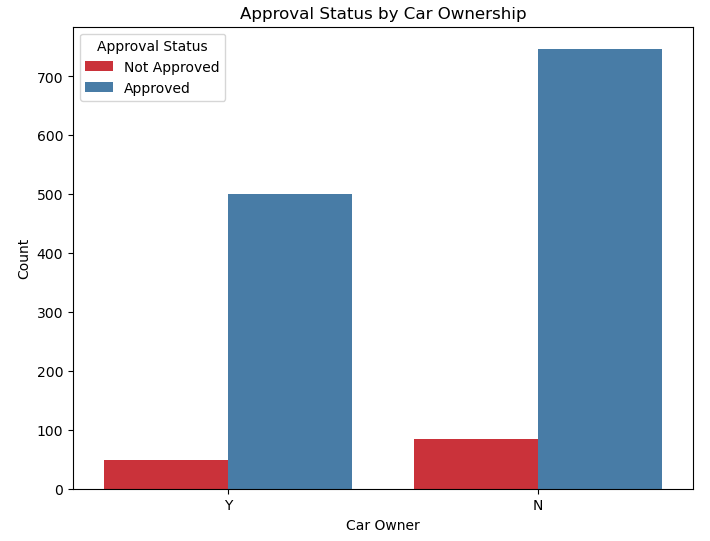
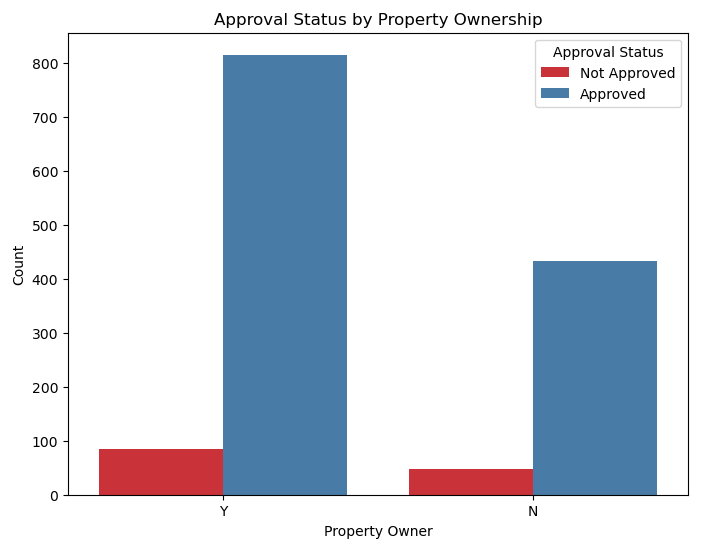
In accordance with my hypothesis, it's plausible to consider that the number of family members can indeed influence credit card approval. Larger families tend to have elevated expenses, potentially resulting in distinct spending patterns that can impact their creditworthiness. This observation underscores the significance of the family size variable in the credit card approval process.

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



Hypothesis has indeed proven accurate: a family size of 2 members often indicates a married couple without children. Furthermore, the data indicates that a larger proportion of married individuals have received credit card approvals. This observation aligns with your initial assumption.

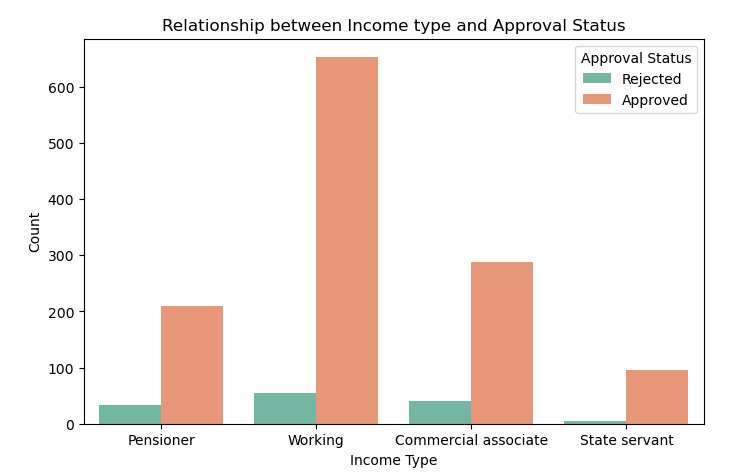
1. Analysing the connection between property and car ownership in relation to approval status.



Based on the provided data, it's evident that the ownership of a car or other vehicle doesn't significantly impact credit card approval, as the count of applicants who don't own a car or vehicle is considerably higher than those who do.

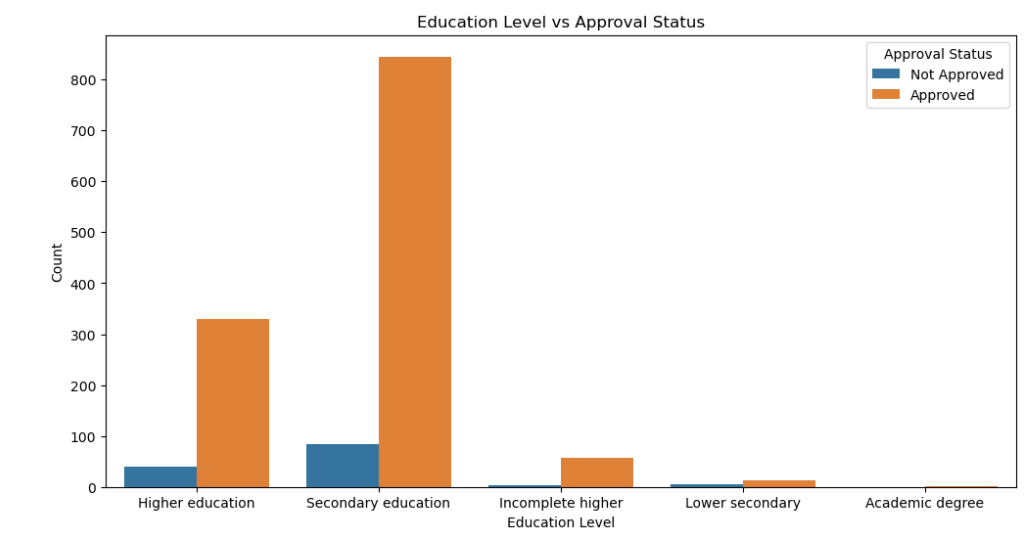
However, property ownership appears to be a more influential factor in credit card approval decisions. Applicants who own property are more likely to receive approval for a credit card. This suggests that property ownership plays a more substantial role in the credit card approval process compared to vehicle ownership.

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



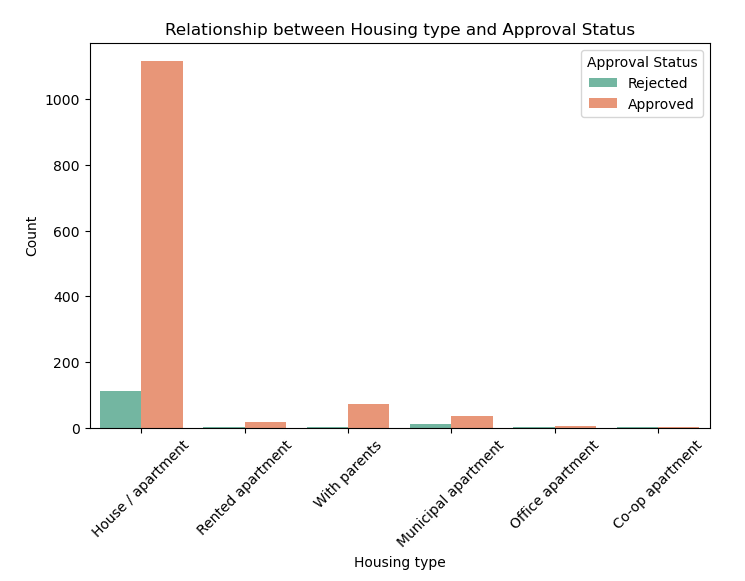
From the data, we can discern a relationship between income type and employment status. It appears that applicants categorized as "working," "commercial associate," and "state servant" are typically considered employed, while applicants labelled as "pensioner" are often categorized as unemployed. This observation suggests a connection between income type and employment status.

1. Analysing the correlation between education and approval status.



The data indicates that the education qualification of the applicant does not significantly affect credit card approval. Most approved applicants have secondary education, while those with a degree have the lowest approval rates. This suggests that education level may not be a significant factor in credit card approval decisions.

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



As observed previously with property ownership, a significant number of applicants possess a house or apartment as a form of property. This reiterates the notion that individuals who own property are more likely to receive credit card approvals.

**Machine Learning Approach**

1. **Method used for machine learning based predictions:**
   1. **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.

* 1. **Data Splitting:**

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

* 1. **Model Selection:**

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

* 1. **Model Training:**

Train the selected models on the training data.

* 1. **Model Evaluation:**

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

* 1. **Model Selection:**

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

1. **Undertaking the essential measures to enhance the accuracy of your model:**

To improve the accuracy of your model, I had taken steps to tune hyperparameters using GridSearchCV for different algorithms.

1. **Logistic Regression**:

Best Hyperparameters: C=10, penalty='l2', solver='liblinear'

1. **Decision Tree:**

Best Hyperparameters: criterion = 'gini', min\_samples\_leaf = 1, min\_samples\_split = 2

1. **Random Forest:**

Best Hyperparameters: criterion = 'gini', min\_samples\_leaf = 1, min\_samples\_split=2, n\_estimators=100, random\_state= 42

1. **Support Vector Classifier (SVC):**

Best Hyperparameters: C=10, gamma=1

1. **Evaluating the optimal model choice by comparing all four available models.:**

To justify the most appropriate model, we need to consider the performance metrics and the specific goals of your credit card approval application. Let's analyse each model:

|  |  |
| --- | --- |
| Model | Accuracy Score |
| Random Forest | 95.26% |
| Support Vector Machine (SVM) | 93.81% |
| Decision Tree | 93.44% |
| Logistic Regression | 60.47% |

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

* **Accuracy Score:** The Random Forest model achieved the highest accuracy score among the models, with a value of 0.95. This indicates that it correctly predicted credit card approvals in 90% of cases.
* **Precision Score:** For predicting approved cases (1), the Random Forest model achieved a precision score of 0.99, signifying that when it predicted an approval, it was highly reliable. For predicting non-approved cases (0), it achieved a precision score of 0.92, indicating a great level of accuracy.
* **Recall Score:** The Random Forest model demonstrated excellent recall for approved cases (1) with a score of 0.91, implying that it effectively captured nearly all actual credit card approvals. While its recall for non-approved cases (0) was lower at 0.99.
* **F1 Score:** The F1 Score, which combines precision and recall, was high for predicting approved cases (1), with a value of 0.95, indicating a good balance between precision and recall. For non-approved cases (0), the F1 Score was 0.96, reflecting a trade-off between precision and recall, as expected.

In conclusion, the Random Forest model demonstrates the best overall performance in terms of accuracy and precision while maintaining a reasonable balance between precision and recall for both approval and non-approval predictions. Therefore, it is the recommended choice for credit card approval prediction based on the provided data and evaluation metrics.