Project Proposal to predict credit card approval:

Title: Predictive Modeling for Credit Card Approval

# Introduction:

The financial industry heavily relies on credit card approval processes to assess the creditworthiness of applicants. Traditional methods involve manual evaluation, which can be time-consuming and prone to errors. Leveraging machine learning algorithms can streamline this process, making it faster, more accurate, and efficient. This project aims to develop a predictive model for credit card approval using historical data and machine learning techniques.

## Objectives:

- Develop a predictive model to automate the credit card approval process.

- Utilize historical credit card application data to train the model.

- Evaluate the model's performance in terms of accuracy, precision, recall, and F1 score.

- Data collection and preprocessing.

- Model selection and training.

- Model evaluation and optimization.

- Development of user interface.

- Model deployment and testing.

- Final adjustments and documentation.

## Expected Outcome:

- A predictive model capable of accurately assessing credit card approval likelihood based on applicant data.

- Improved efficiency and accuracy compared to manual approval processes.

- A user-friendly interface for easy application of the model in real-world scenarios.

- Documentation detailing the model's development, implementation, and usage guidelines.

## Conclusion:

Automating credit card approval processes through predictive modeling can offer significant advantages in terms of efficiency, accuracy, and customer satisfaction. By leveraging historical data and machine learning techniques, this project aims to develop a robust model that can streamline the approval process while maintaining compliance with regulatory standards.

## Explanation:

The hypothesis for data analysis aims to investigate whether various factors such as demographic information (age, gender, marital status), financial history (credit score, income), and employment status (employment type, years of employment) have a statistically significant impact on credit card approval decisions. The null hypothesis states that these factors do not influence credit card approval, while the alternative hypothesis suggests that they do. Through data analysis and statistical testing, we will seek to validate or reject these hypotheses, providing insights into the key determinants of credit card approval.

# SECTION 1

1. In today's world, the proposal for credit card approval is crucial due to several reasons. Firstly, credit cards play a significant role in the modern economy, facilitating transactions and enabling consumers to access goods and services conveniently. As such, the approval process ensures that credit is extended to individuals who are likely to use it responsibly, reducing the risk of default and financial losses for the issuing bank. Predicting a good client is valuable for a bank because it allows them to assess the creditworthiness of applicants accurately. By analyzing factors such as credit history, income level, payment behavior, and debt-to-income ratio, banks can make informed decisions about extending credit to clients who are more likely to repay their debts on time, thus minimizing the risk of defaults and improving overall financial performance. Predictive modeling and data analysis techniques are often used to evaluate these factors and predict the likelihood of a client being a good credit risk.
2. The impact of credit card approval proposals and the ability to predict good clients extends significantly to the banking sector. Firstly, it helps banks manage risk more effectively by identifying and approving clients who are less likely to default on their credit card payments. This reduces the number of non-performing assets and improves the overall asset quality of banks, enhancing their financial stability and resilience. Additionally, by targeting the right clients for credit card offerings, banks can optimize their marketing strategies and customer acquisition efforts, leading to higher profitability and customer satisfaction. Moreover, accurate prediction of good clients enables banks to tailor their credit products and services more effectively, offering competitive interest rates, rewards, and benefits to attract and retain valuable customers. Overall, these practices contribute to a healthier and more sustainable banking sector with improved risk management, profitability, and customer relationships.
3. One potential gap in knowledge regarding credit card approval and client prediction in the banking sector in India could be related to the specific nuances of the Indian market, including cultural and socioeconomic factors that may impact credit behavior differently compared to other regions. Understanding these nuances and incorporating them into predictive models and decision-making processes can be crucial for banks operating in India. To address this gap, a proposed method could involve leveraging advanced data analytics and machine learning algorithms tailored to the Indian market. This could include collecting and analyzing a wide range of data points such as spending patterns, demographic information, credit histories, employment status, and regional economic indicators. By developing models that take into account these factors unique to India, banks can improve the accuracy of predicting good clients and making informed credit card approval decisions. Furthermore, continuous monitoring and updating of these predictive models based on evolving market trends and regulatory changes can ensure their relevance and effectiveness over time. Collaborating with data scientists, industry experts, and regulatory authorities can also provide valuable insights and guidance in refining these methods to meet the specific needs and challenges of the banking sector in India. Overall, leveraging advanced data analytics and tailored predictive modeling can help Indian banks enhance risk management, customer targeting, and overall performance in the credit card approval process.

---x---

# SECTION 2

## Hypothesis:

Hypothesis for Data Analysis:

Null Hypothesis (H0): There is no significant relationship between the applicant's demographic information, financial history, and employment status, and the likelihood of credit card approval.

Alternative Hypothesis (H1): There exists a significant relationship between the applicant's demographic information, financial history, and employment status, and the likelihood of credit card approval.

Initial Hypothesis:

Based on the questions we want to address on data analysis (DA) track, we hypothesize that certain key features will significantly impact a machine learning (ML) model's ability to predict credit card approval outcomes in the banking sector. Our assumptions include:

* **Credit History:** We assume that a positive credit history with no or minimal defaults will strongly correlate with higher chances of credit card approval.
* **Income Level:** Higher income levels are expected to be positively associated with credit card approval, as they indicate a client's ability to repay debts.
* **Debt-to-Income Ratio:** A lower debt-to-income ratio is hypothesized to be beneficial, indicating a client's manageable debt burden relative to income.
* **Payment Behavior:** Clients with a history of timely payments and low delinquency rates are expected to have better chances of credit card approval.
* **Employment Status:** Stable employment and regular income are assumed to be positive factors influencing credit card approval decisions.
* **Demographic Factors:** Factors such as age, gender, and location may also impact credit card approval rates, with younger demographics and urban areas potentially showing higher approval rates.

Our initial hypothesis suggests that these features will exhibit discernible patterns in the data and will play a crucial role in developing an effective ML model for predicting credit card approval outcomes in the banking sector.

----x----

# SECTION 3:

Refer creditfile\_py

## Q1

To prove or disprove the initial hypothesis regarding the impact of key features on credit card approval outcomes in the banking sector, we would adopt the following approach:

1. Data Collection: credit\_card.csv
2. Data Preprocessing: Cleanse the data by handling missing values, outliers, and inconsistencies. Normalize or standardize numerical features, encode categorical variables, and split the data into training and testing sets.
3. Exploratory Data Analysis (EDA): Conduct EDA to analyze the relationships between the features and the target variable (credit card approval). Explore statistical summaries, visualizations (e.g., histograms, scatter plots, correlation matrices), and identify potential patterns or trends in the data.
4. Feature Selection: Use techniques like correlation analysis, feature importance ranking from ML algorithms (e.g., Random Forest), and domain knowledge to select the most relevant features that impact credit card approval outcomes.
5. Model Building: Develop ML models (e.g., logistic regression, decision trees, random forest, gradient boosting) to predict credit card approval based on the selected features. Train the

accuracy, precision, recall, and F1-score.

1. Hypothesis Testing: Conduct hypothesis testing using statistical methods (e.g., t-tests, ANOVA) or hypothesis validation within the context of ML model evaluation. Evaluate whether the observed patterns and relationships in the data support or refute the initial hypothesis regarding the impact of key features on credit card approval outcomes.
2. Validation and Interpretation: Validate the ML models using the testing dataset to assess their generalization performance. Interpret the model results, feature importance scores, and statistical significance to draw conclusions about the hypothesized impact of features on credit card approval.

By following this approach, we can systematically analyze the data, build predictive models, and evaluate the hypothesis regarding the influence of key features on credit card approval outcomes in the banking sector.

## Q2

For the project focused on predicting credit card approval outcomes in the banking sector based on key features, several feature engineering techniques would be relevant to enhance the predictive power of the machine learning models. These techniques include:

* **One-Hot Encoding:** Convert categorical variables (such as employment status, gender, location) into binary vectors to represent each category as a separate feature. This allows the ML models to handle categorical data effectively.
* **Feature Scaling:** Normalize numerical features like income level and debt-to-income ratio to ensure that all features are on a similar scale. Common scaling techniques include Min-Max scaling and Standardization (Z-score normalization).
* **Handling Missing Values:** Impute missing values in the dataset using methods such as mean/median imputation, mode imputation for categorical variables, or advanced techniques like K-Nearest Neighbors (KNN) imputation.
* **Feature Transformation:** Apply transformations such as logarithmic transformation or square root transformation to numerical features to make their distributions more Gaussian-like, which can improve the performance of certain ML algorithms.
* **Interaction Features:** Create new features by combining existing features through mathematical operations (e.g., product, ratio, sum) to capture potential interactions and nonlinear relationships between variables.

By applying these feature engineering techniques, we can preprocess and transform the dataset to create informative and predictive features that improve the accuracy and robustness of the ML models for predicting credit card approval outcomes in the banking sector.

## Q3

The data analysis process proposed for predicting credit card approval outcomes in the banking sector is justified by its systematic and comprehensive approach. Beginning with data collection and preprocessing ensures that the dataset is clean, relevant, and ready for analysis. The subsequent step of exploratory data analysis (EDA) provides valuable insights into the relationships between features and the target variable, aiding in feature selection and hypothesis formulation. By employing feature selection techniques, the most relevant features impacting credit card approval can be identified, reducing noise and improving model performance. Model building and evaluation allow for the comparison of different machine learning algorithms, ensuring the selection of the most effective model for prediction. Hypothesis testing adds a layer of statistical validation to the analysis, ensuring the robustness of the conclusions drawn. Finally, interpretation of the model results and insights derived provide actionable information for decision-making within the banking sector, such as optimizing credit risk assessment strategies and customer targeting approaches. Overall, this data analysis process is designed to be rigorous, transparent, and aligned with the project's objectives, making it a justified and effective approach for predicting credit card approval outcomes in the banking industry.

## Q4

Income Distribution:

The histogram of the 'income' feature shows that the distribution is skewed to the right, indicating that a majority of individuals have lower incomes.

However, individuals with higher incomes seem to have a slightly higher likelihood of credit card approval based on the boxplot.

Age Distribution:

The histogram of the 'age' feature suggests a relatively even distribution across different age groups.

There doesn't seem to be a clear trend between age and credit card approval based on the boxplot.

Categorical Features:

The count plots of categorical features like 'employment\_status' and 'marital\_status' show varying distributions among different categories.

For example, employed individuals seem to have a higher likelihood of credit card approval compared to unemployed or self-employed individuals.

Correlation Matrix:

The correlation matrix provides insights into the pairwise correlations between numerical features.

However, the correlation between individual features and the target variable approval status is not explicitly visible in the correlation matrix.

----x----

# SECTION 4

1

For a machine learning task based on credit card approval, we need to predict whether an applicant will be approved for a credit card or not. This is a binary classification problem. Several machine learning algorithms can be suitable for this task. The choice of algorithm depends on various factors including the size of the dataset, the complexity of the data, interpretability requirements, and computational resources. Here are some commonly used algorithms for binary classification tasks like credit card approval:

Naive Bayes:

Naive Bayes classifiers are probabilistic classifiers based on Bayes' theorem with the "naive" assumption of independence between features.

They are simple, fast, and perform well on datasets with a large number of features.

Naive Bayes classifiers are particularly useful when dealing with text classification tasks, but they can also be applied to other types of data.

While Naive Bayes classifiers may not capture complex relationships between features, they often provide competitive performance, especially on smaller datasets.

Logistic Regression:

Logistic regression is a simple and interpretable algorithm suitable for binary classification tasks.

It works well when the relationship between features and the target variable is approximately linear.

Logistic regression also provides probabilities for class membership, which can be useful for decision-making.

Decision Trees:

Decision trees are versatile and easy to interpret.

They can capture complex nonlinear relationships between features and the target variable.

However, decision trees tend to overfit the training data, which can be mitigated using techniques like pruning or ensemble methods.

Random Forest:

Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions.

It typically performs well on a wide range of datasets and is less prone to overfitting compared to individual decision trees.

Random Forest can handle high-dimensional data and provides feature importances, which can be useful for feature selection.

Support Vector Machines (SVM):

SVMs aim to find the hyperplane that best separates the classes in the feature space.

They can handle high-dimensional data and are effective in cases where the decision boundary is nonlinear.

SVMs are memory-efficient, especially when using a kernel trick to map data into higher-dimensional space.

2

Justification:

Random Forest emerges as the most fitting model for the credit card approval prediction task due to its robust performance and versatility. In the context of this problem, where accuracy and generalization are crucial, Random Forest offers several advantages. Firstly, it effectively handles high-dimensional data and maintains predictive accuracy even with a large number of features, making it suitable for datasets with diverse and potentially correlated attributes. Secondly, Random Forest mitigates overfitting by constructing multiple decision trees on random subsets of the data and combining their predictions through ensemble learning, thereby reducing the variance of the model and enhancing its ability to generalize to unseen data. Additionally, Random Forest provides insights into feature importance, aiding in feature selection and understanding the underlying factors influencing credit card approval decisions. Overall, the ensemble nature of Random Forest, coupled with its ability to handle complex datasets and mitigate overfitting, positions it as the most appropriate model for accurately predicting credit card approval while maintaining robust performance on diverse datasets.

Other than that, Considering the task of credit card approval prediction, where interpretability and performance are both important, a model that balances these factors would be ideal. Logistic Regression stands out as a strong candidate due to its simplicity, interpretability, and efficiency. It provides probabilities for class membership, allowing for easy interpretation of results and decision-making. Additionally, Logistic Regression is less prone to overfitting compared to more complex models like neural networks or gradient boosting machines, which is advantageous given the potential regulatory and compliance requirements in the financial industry.

However, it's important to validate the performance of Logistic Regression against other models through rigorous experimentation and cross-validation. If Logistic Regression doesn't meet the desired performance criteria, ensemble methods like Random Forest or SV Machines could be considered as they offer a good balance between performance and interpretability. Ultimately, the choice of the most appropriate model should be based on empirical evidence from thorough evaluation on the specific dataset and requirements of the credit card approval prediction task.

3

Grid searching is a method to find the best possible combination of hyper-parameters at which the model achieves the highest accuracy. REFER CODE.

4

Naive Bayes:

Naive Bayes achieved the lowest accuracy of approximately 14%, indicating poor performance.

While it showed high recall for class 1 (defaulters), its precision for class 1 was very low.

The F1-score for class 1 was also low, indicating poor balance between precision and recall.

Logistic Regression:

Logistic Regression achieved an accuracy of 90%, which is significantly higher than Naive Bayes.

It showed higher precision, recall, and F1-score for class 0 (non-defaulters) compared to Naive Bayes.

However, its performance for class 1 (defaulters) was still relatively poor.

Support Vector Machine (SVM):

SVM achieved an accuracy of approximately 91%, which is slightly higher than Logistic Regression.

It showed high precision and recall for class 0, similar to Logistic Regression.

However, its performance for class 1 was better than Logistic Regression, with higher precision, recall, and F1-score.

Random Forest:

Random Forest achieved the highest accuracy of approximately 94% among all models.

It showed excellent precision, recall, and F1-score for class 0, indicating strong performance in identifying non-defaulters.

For class 1, Random Forest achieved perfect precision and relatively high recall and F1-score, indicating effective identification of defaulters.

Decision Tree:

Decision Tree achieved an accuracy of approximately 89%.

It showed good precision, recall, and F1-score for class 0, similar to Logistic Regression and SVM.

However, its performance for class 1 was weaker compared to Random Forest, with lower precision, recall, and F1-score.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision\_Class0 | Precision\_Class1 | Recall\_Class0 | Recall\_Class1 | F1-score\_Class0 | F1-score\_Class1 |
| Naive Bayes | 0.1387 | 1 | 0.1 | 0.05 | 1 | 0.09 | 0.18 |
| Logistic Regression | 0.9 | 0.91 | 0.33 | 0.99 | 0.03 | 0.95 | 0.06 |
| Support Vector Machine | 0.9129 | 0.91 | 0.8 | 1 | 0.13 | 0.95 | 0.23 |
| Random Forest | 0.9355 | 0.93 | 1 | 1 | 0.33 | 0.97 | 0.5 |
| Decision Tree | 0.8871 | 0.92 | 0.37 | 0.96 | 0.23 | 0.94 | 0.29 |

Overall, Random Forest appears to be the best-performing model among the evaluated models, as it achieved the highest accuracy and showed strong performance in identifying both non-defaulters and defaulters. However, further analysis, including feature importance and model tuning, may be necessary to make a final decision on the best model for the task.

-------X------

# **MY SQL queries**

**THE DATASET IS CLEANED USING PANDAS IN PYTHON. REFER CODE FILE 2**

create database credit;

use credit;

show databases;

--this table is imprted from cleaned\_credit\_card.csv file using mysql workbence

SELECT \* FROM cleaned\_credit\_card

-- Group the customers based on their income type and find the average of their annual income.

SELECT Type\_Income, AVG(Annual\_income) AS average\_income

FROM cleaned\_credit\_card

GROUP BY Type\_Income;

-- Find the female owners of cars and property.

SELECT \*

FROM cleaned\_credit\_card

WHERE GENDER = 'F'

AND (Car\_Owner = 'Y' OR Propert\_Owner = 'Y');

-- Find the male customers who are staying with their families.

SELECT \*

FROM cleaned\_credit\_card

WHERE GENDER = 'M'

AND Housing\_type = 'With parents ' or Family\_Members > 0;

-- Please list the top five people having the highest income.

SELECT \*

FROM cleaned\_credit\_card

ORDER BY Annual\_income DESC

LIMIT 5;

-- What is the highest education level and what is the total count?

SELECT EDUCATION , COUNT(\*) AS higher\_education\_count

FROM cleaned\_credit\_card

WHERE EDUCATION = 'Higher education';

-- How many married people are having bad credit?

SELECT COUNT(\*) AS married\_bad\_credit\_count

FROM cleaned\_credit\_card

WHERE Marital\_status = 'Married'

AND label = 1;

-- Between married males and females, who is having more bad credit?

SELECT GENDER, COUNT(\*) AS bad\_credit\_count

FROM cleaned\_credit\_card

WHERE Marital\_status = 'Married'

AND label = 1

GROUP BY GENDER;