**Capstone Project 1 : Credit card**

**Section 1: Questions**

1.Why is your proposal important in today’s world?

Predicting credit card approval accurately is crucial for banks to minimize risk and enhance profitability. Accurate predictions help in identifying good clients, reducing the chance of defaults, and improving customer satisfaction. This proposal aims to enhance the credit card approval process, leading to better financial decisions and more efficient risk management.

2.How is it going to impact the banking sector?

Implementing an advanced predictive model can streamline the credit approval process, reduce manual errors, and ensure that approvals and rejections are based on data-driven insights. This can lead to more personalized financial services and improved customer experience.

3.What is the gap in the knowledge or how can your method be helpful for any bank in India?

Current models might not fully leverage new data sources or advanced analytical techniques. This proposal addresses this gap by integrating machine learning techniques and external data to provide more accurate and reliable predictions, especially relevant in the rapidly evolving financial sector in India.

**Section 2: Initial Hypothesis (or hypotheses)**

* Income and Approval Hypothesis: Higher annual income correlates with a higher likelihood of credit card approval. (Test using regression analysis)
* Property Ownership Hypothesis: Owning property is positively associated with credit card approval. (Test using correlation analysis)
* Employment Status Hypothesis: Employed individuals have a higher probability of credit card approval compared to unemployed individuals. (Test using t-tests or logistic regression)
* Contact Information Hypothesis: Multiple contact methods (e.g., mobile phone, work phone) increase the likelihood of approval. (Test using feature importance analysis)

**Section 3: Data analysis approach**

1. What approach are you going to take in order to prove or disprove your hypothesis?

Conduct Exploratory Data Analysis (EDA) to identify key patterns and relationships in the data. For example, examine distributions of income, employment status, and demographic features.

1. What feature engineering techniques will be relevant to your project?

* Categorical Encoding: Convert categorical variables (e.g., gender, marital status) into numerical formats.
* Handling Missing Values: Impute or remove missing data as appropriate.
* Normalization: Scale features to ensure consistent input ranges for machine learning models.

1. Please justify your data analysis approach?

EDA helps in understanding data patterns and relationships, which informs feature selection and model building. Effective feature engineering ensures that the data fed into the model is relevant and useful for prediction.

1. Identify important patterns in your data using the EDA approach to justify your findings?

Use visualizations and correlation matrices to identify key patterns and relationships between features and approval rates.

**Section 4: Machine learning approach**

1. What method will you use for machine learning based predictions for credit card approval?
2. Logistic Regression: Good for binary classification and interpretable results.
3. Random Forest: Handles feature importance and interactions well.
4. Gradient Boosting Machines (GBM): Provides high accuracy and handles complex relationships.
5. Support Vector Machines (SVM): Effective in high-dimensional spaces.
6. Please justify the most appropriate model.
7. Logistic Regression is chosen for its interpretability and efficiency in handling binary outcomes. It provides probabilities for each class and helps in understanding feature impact.
8. Please perform necessary steps required to improve the accuracy of your model.
9. Hyperparameter Tuning: Optimize model parameters using techniques like grid search or random search.
10. Cross-Validation: Use k-fold cross-validation to assess model performance and generalizability.
11. Feature Selection: Refine features based on importance scores and model performance.
12. Please compare all models (at least 4 models).
13. Logistic Regression: Simple and interpretable but may lack accuracy with complex data.
14. Random Forest: High accuracy and handles feature importance well.
15. Gradient Boosting Machines: Offers high performance with complex datasets.
16. Support Vector Machines: Effective for high-dimensional data but may be computationally intensive.

**SQL:**

Notebook Link:

[Colab Notebooks - Google Drive](https://drive.google.com/drive/folders/1_8Oqn6AKIwfqCySHtHRtVRI2wodQHmAC)