**Section 1:**

1. **Why is your proposal important in today’s world? How predicting a good client is worthy for a bank?**

In today's world banks operate in changing financial environments, such as global crises and financial market fluctuations.So risk management is very crucial in maintaining bank stability. Efficiently predicting the clients worthiness is an important factor, so that the banks can optimise their profitability, as credit plays an important role in banks stability and profitability

1. **How is it going to impact the banking sector?**

Risks like financial losses, loan defaults can be reduced efficiently. Recognising good clients can help to give better customer service which would be a loyal remark for the clients, as loyal customers are important assets to a bank.

1. **If any, what is the gap in the knowledge or how your proposed method can be helpful if required in future for any bank in India.**

Mentioning a gap as a client's existing debt, debt-to-income ratio is a crucial factor in accessing a client's credit worthiness.

As always, lending credit is the main source of profitability of any bank, even in the future. The proposed method can predict the trustworthy clients, so that banks have an efficient risk management system, which can reduce loan debts and future financial instability.

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

1. **If DA track please aim to identify patterns in the data and important features that may impact a ML model?**

The labels are numerical so correlation can be used to identify the pattern within the features.

1. **If ML track please perform part ‘i’ as well as multiple machine learning models, perform all required steps to check if there is any assumption and justify your model. Why is your model better than any other possible model? Please justify it by relevant cost functions and if possible by any graph.**

Logistic Regression, Decision Tree, Support Vector Machine and Random Forest models are used for comparision. Within these models Random Forest performs better than all the other models specified above.

**Section 3: Data analysis approach.**

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

Univariant analysis can be useful to find out the data dispersion and missing values within the data and bivarient analysis for finding the pattern of feature variables with respect to target variables.

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

Missing Values: Impute missing values in features like mobile\_phone, work\_phone, email\_id, and family\_members using methods like mean, median, or mode for categorical features and median or KNN imputation for numerical features.

* Outliers: Identify and handle outliers in features like annual\_income, employed\_days, and birthday\_count using methods like winsorization or clipping.
* Feature Scaling: Scale numerical features like annual\_income, employed\_days, and family\_members to a common range (e.g., between 0 and 1) using standardization or normalization techniques.

1. **Please justify your data analysis approach.**

Focus on Relevant Features:

* Children: Having children could impact financial stability and potentially affect card approval, hence the binary feature indicating their presence.
* Annual Income: Income is a crucial indicator of affordability and repayment potential, justifying income-related features and categorization.
* Age: Age reflects financial independence and stability, hence features based on birthday\_count and derived age groups.
* Employment History: Employment duration signifies stability and income source, prompting features based on employed\_days and categorization.
* Contact Information: Phone numbers and email validity can indicate responsible behavior and communication channels.
* Family Size: Family size affects financial obligations and risk, warranting family size features and categorization.

Feature Engineering Techniques:

* Data Cleaning and Preprocessing: Missing values, outliers, and scaling address data quality issues, improving model accuracy.
* Interaction and Derived Features: Combining existing features creates richer representations, potentially revealing hidden patterns.
* Feature Selection: Removing irrelevant or redundant features reduces model complexity and improves interpretability.

Overall, the proposed data analysis approach aims to create a robust and interpretable model for predicting credit card approval by emphasizing relevant features, applying appropriate techniques, and considering ethical aspects. This approach provides a solid foundation for further refinement and customization based on specific data and objectives.

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

* annual\_income which is a sensible variable w.r.t the target variable is right skewed.
* type\_ocupation column having more than 30% missing values so can be dropped.
* There are features with different scale like in birthday\_count and employed\_dates that can be scaled

**Section 4: Machine learning approach**

1. **What method will you use for machine learning based predictions for credit card approval?**

The problem statement is classification based , so models like logistic regression, random forest classifier can be suitable for better prediction.

1. **Please justify the most appropriate model.**

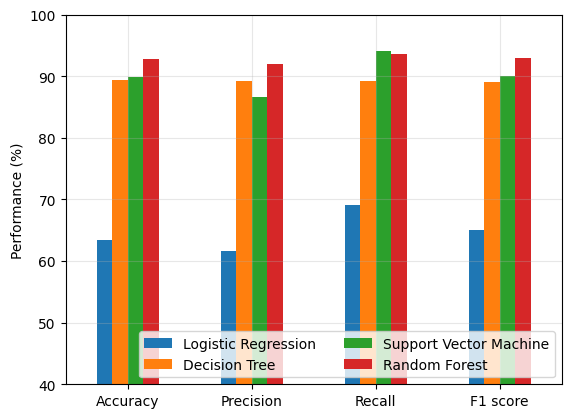
Random forest can be a suitable model for classification as it combines the predictions of multiple decision trees, achieving higher accuracy and robustness compared to individual trees. This "wisdom of the crowds" approach reduces the risk of overfitting and improves generalizability to unseen data.

1. **Please perform necessary steps required to improve the accuracy of your model.**

Hyper parameter tuning is performed to increase the accuracy. Parameters like estimators, maximum features, maximum depth of tree etc. are considered.

1. **Please compare all models (at least 4 models).**

From all of the four models, the random forest model performs better to the problem statement.



The above graph, clearly shows Random Forest performing better than the other models.