Capstone Project (Credit Card Approval)

SUBIR BHAKAT (S3607)

---

1. **Why is your proposal important in today's world?**
    - **Enhances Financial Inclusion:** Improving the credit approval process makes credit card access more inclusive.
    - **Embraces Data-Driven Decision-Making:** Aligns with the trend of using data and technology to enhance financial services.
    - **Mitigates Financial Risks:** Helps financial institutions make accurate credit decisions, reducing default risks.
    - **Boosts Efficiency and Cuts Costs:** Automates processes, leading to cost savings in a competitive financial landscape.
    - **Personalizes Services:** Tailor's credit approval decisions to individual financial profiles.
    - **Ensures Regulatory Compliance:** Aligns with strict financial regulations, promoting responsible lending.
    - **Enhances Fraud Detection:** Improves security by detecting fraudulent applications.
    - **Prioritizes Data Privacy and Security:** Safeguards sensitive customer data while making informed credit decisions.

    How predicting a good client is worthy for a bank?

    - **Reduces Default Risk:** Lowers the risk of loans not being repaid, safeguarding the bank's financial stability.
    - **Lower Costs:** Decreases provisions for loan losses and collection expenses, saving money.
    - **Enhances Portfolio Quality:** Improves asset quality and credit ratings, attracting investors.
    - **Expands Lending Capacity:** Allows for more lending, potentially attracting more clients and revenue.
    - **Retains Customers:** Keeps responsible clients, fostering long-term relationships.
    - **Competitive Edge:** Offers better terms, gaining a competitive advantage.
    - **Ensures Compliance:** Helps in responsible lending, avoiding legal issues and fines.
    - **Informs Data-Driven Decisions:** Supports better decision-making in a data-centric world.

2. **How is it going to impact the banking sector?**
    - **Reduced Defaults:** A more accurate prediction of creditworthiness leads to fewer loan defaults, which, in turn, strengthens the overall financial health of banks.
    - **Improved Profitability:** Lower defaults and reduced loan loss provisions boost a bank's profitability, potentially leading to better financial performance.
    - **Enhanced Risk Management:** Banks can more effectively manage risk by identifying and addressing potential defaulters early in the lending process.
    - **Stability:** Fewer defaults contribute to a more stable banking sector, reducing the likelihood of financial crises.
    - **Competitive Advantage:** Banks with strong predictive models can offer more attractive loan terms, gaining a competitive edge in the market.
    - **Regulatory Compliance:** Predictive models help banks comply with regulatory requirements, avoiding penalties and legal issues.
    - **Innovation:** By harnessing data and analytics, banks can continuously innovate and adapt to changing market conditions and customer behaviors.

- **Customer Trust:** Accurate credit decisions build trust with customers, fostering long-term relationships and brand loyalty.
- **Financial Inclusion:** Improved credit assessment can expand financial inclusion by extending credit to underserved populations.
- **Investor Confidence:** A reduced risk of loan defaults can attract investors and enhance the bank's credit rating.

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

   Knowledge Gap:

   - **Local Variations:** India's economic, regulatory, and cultural factors create unique challenges for banks.
   - **Impact on Credit Risk:** Local spending behavior, income distribution, and regional variations can affect credit risk.
   - **Data Needs:** Addressing these nuances requires access to detailed, localized data.

   How my proposed method can be helpful:

   - **Customization:** My method can be customized to consider local data patterns, ensuring that credit decisions align with the specific needs and risk factors in the Indian market.
   - **Adaptability:** As the banking landscape evolves in India, my method can adapt to changing economic conditions and regulatory requirements. Machine learning models can continuously learn from new data, allowing banks to stay up-to-date with emerging trends.
   - **Scalability:** My method can efficiently process and evaluate a vast number of credit applications, especially valuable in a country with a sizable population like India.
   - **Responsible Lending:** Banks in India are increasingly focused on responsible lending practices to mitigate risk and comply with regulations. My method can help banks make responsible lending decisions by identifying clients with a lower risk of default.
   - **Financial Inclusion:** India's push for financial inclusion means that many individuals and businesses are seeking access to credit for the first time. My method can contribute to expanding financial services to underserved populations while managing associated risks effectively.
   - **Competitive Edge:** Banks that adopt advanced predictive methods gain a competitive edge by offering tailored credit solutions, attracting more clients, and building a strong reputation for responsible lending.

Initial Hypothesis (or hypotheses)

1. **Here you have to make some assumptions based on the questions you want to address based on the DA track or ML track.**

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

      Assumptions:

      - The dataset contains information related to credit card applications, including applicant demographics, financial information, and application details.
      - The goal is to perform data analysis to understand the dataset's characteristics and relationships among variables before building a predictive model.

      Data Analysis and Feature Identification:

      - There are some important features that may impact our ML model. Those are Propert_Owner, Annual_income, Type_Income, Housing_type, Employed_days, and Family_Members.

   b. **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.**

      - It is a Classification problem so, we can’t use the Linear Regression model to predict the output. Classification problem having the value ‘yes’ and ‘no’ or ‘0’ and ‘1’. If we use the Linear regression model we can’t reach the accuracy level. So, for a Classification problem, we can use Logistic Regression or any tree-based model such as a Decision Tree or Random Forest.
      - We may need to Tuning the ML models for more accurate results. Here we are going to use Logistic Regression, a Decision Tree, and Random Forest ML models and at the end, we will compare them to find the best ML model for this problem.

Data Analysis Approach:

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

- **Data Exploration:** Begin by loading and exploring the dataset. Check the structure, data types, and summary statistics for each variable. There are two CSV files,
- **Credit_card.csv contains:** Ind_ID, GENDER, Car_Owner, Propert_Owner, CHILDREN, Annual_income, Type_Income, EDUCATION, Marital_status, Housing_type, Birthday_count, Employed_days, Mobile_phone, Work_Phone, Phone, EMAIL_ID, Type_Occupation, and Family_Members
- **Credit_card_label.csv contains:** Ind_ID, and label
- To perform further analysis these two csv files must merge.

2. What feature engineering techniques will be relevant to your project? 3. Please justify your data analysis approach. 4. Identify important patterns in your data using the EDA approach to justify your findings.

- **Data Cleaning:** Address missing values, duplicates, and outliers. Cleaning the data ensures that subsequent analysis is based on reliable information.
    - **Filling NaN values:** After a quick view of the merged dataset, it is found that there are 1548 entries and 19 columns. Replacing all NaN values from The 'GENDER', 'Annual_income', and 'Birthday_count' columns.
    - **Deleting column:** The 'Type_Occupation' column has more than 30% NaN values. If we replace all the NaN values in this column then the predicted values will be incorrect. So, deleting the whole column is the best move. Also, this column is not very important for the ML model.
    - **Finding the Outliers:** Using the 'describe()' function we can take a look at all the data and understand if there are any Outliers present or not. Also, we use the 'kdeplot' to visualize the outliers in a graphical form. After the analysis, we found some Outliers in the 'Employed_days', and 'Annual_income' columns. Those Outliers are removed using IQR.

      After performing these processes the remaining rows are 1121 and the columns are 18.

- **Data Encoding:** To make the data suitable for the ML model we need to encode all categorical columns into numerical columns. So, we do 'Dummy Encoding' for the 'GENDER', 'Type_Income', 'Marital_status', and 'Housing_type' columns. 'Binary Encoding' for the 'Car_Owner', and 'Propert_Owner' columns. 'Ordinal Encoding' for the 'EDUCATION' column.

  After Encoding the remaining rows are 1121 and the columns are 27.

Machine Learning Approach:

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

    - **Model Selection:** I have Choosen multiple machine learning algorithms suitable for binary classification tasks like logistic regression, decision trees, and random forest.

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

    - **Model Justification:** Justify why a particular model is chosen as the best option based on the following factors:
        - **Performance Metrics:** Select the model that yields the highest values for relevant performance metrics, such as accuracy, precision, recall, or F1-score. Justify the choice based on the specific goals of the credit card approval prediction task.
        - **Interpretability:** Consider the interpretability of the model. If interpretability is a crucial requirement, simpler models like logistic regression or decision trees might be preferred.
        - **Computational Resources:** Assess the computational resources required to train and deploy the model. Choose a model that balances computational efficiency with performance.
        - **Model Complexity:** Evaluate the model's complexity and the risk of overfitting. Simpler models may be preferred if there's limited data or a need for transparency.
        - **Robustness:** Consider the model's robustness to variations in the dataset or real-world scenarios.

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

    - **Model Training:** Train each selected model using the training dataset.
    - **Model Evaluation:** Evaluate each model's performance on the testing dataset using appropriate evaluation metrics like accuracy, precision, L1-score, solver, max_depth and max_features. Assess how well the models predict credit card approval.

4. **Please compare all models.**

    - **Model Comparison:** The accuracy score of Logistic Regression was around 90%, and after tuning the accuracy was remain same.
      The accuracy score of Decision tree was around 87%, and after tuning the accuracy was became around 90%.
      The accuracy score of Random forest was around 90%, and after tuning the accuracy was remain same.
      After tuning all models has the same accuracy score. But when we check the accuracy score on the Training data, we saw the accuracy score for Random forest was around 95% , for Decision Tree 99% , and for Logistic Regression 92%. So, we can say if all models prediction for output results are same but for training data the accuracy score is different then we must choose the model which has better accuracy score in training data. Here, I am choosing Decision Tree model because it’s accuracy score for training data was 99%.