**Feature Selection, Model Selection and Tuning**

**Background & Context**

The Thera bank recently saw a steep decline in the number of users of their credit card, credit cards are a good source of income for banks because of different kinds of fees charged by the banks like annual fees, balance transfer fees, and cash advance fees, late payment fees, foreign transaction fees, and others. Some fees are charged on every user irrespective of usage, while others are charged under specified circumstances.

Customers’ leaving credit cards services would lead bank to loss, so the bank wants to analyze the data of customers’ and identify the customers who will leave their credit card services and reason for same – so that bank could improve upon those areas

You as a Data scientist at Thera bank need to come up with a classification model that will help bank improve their services so that customers do not renounce their credit cards

**Objective**

1. Explore and visualize the dataset.
2. Build a classification model to predict if the customer is going to churn or not
3. Optimize the model using appropriate techniques
4. Generate a set of insights and recommendations that will help the bank

**Data Dictionary:**  
\* CLIENTNUM: Client number. Unique identifier for the customer holding the account  
\* Attrition\_Flag: Internal event (customer activity) variable - if the account is closed then 1 else 0  
\* Customer\_Age: Age in Years  
\* Gender: Gender of the account holder  
\* Dependent\_count: Number of dependents  
\* Education\_Level: Educational Qualification of the account holder  
\* Marital\_Status: Marital Status of the account holder  
\* Income\_Category: Annual Income Category of the account holder  
\* Card\_Category: Type of Card  
\* Months\_on\_book: Period of relationship with the bank  
\* Total\_Relationship\_Count: Total no. of products held by the customer  
\* Months\_Inactive\_12\_mon: No. of months inactive in the last 12 months  
\* Contacts\_Count\_12\_mon: No. of Contacts in the last 12 months  
\* Credit\_Limit: Credit Limit on the Credit Card  
\* Total\_Revolving\_Bal: Total Revolving Balance on the Credit Card  
\* Avg\_Open\_To\_Buy: Open to Buy Credit Line (Average of last 12 months)  
\* Total\_Amt\_Chng\_Q4\_Q1: Change in Transaction Amount (Q4 over Q1)  
\* Total\_Trans\_Amt: Total Transaction Amount (Last 12 months)  
\* Total\_Trans\_Ct: Total Transaction Count (Last 12 months)  
\* Total\_Ct\_Chng\_Q4\_Q1: Change in Transaction Count (Q4 over Q1)  
\* Avg\_Utilization\_Ratio: Average Card Utilization Ratio

**Best Practices for Notebook :**

* The notebook should be well-documented, with inline comments explaining the functionality of code and markdown cells containing comments on the observations and insights.
* The notebook should be run from start to finish in a sequential manner before submission.
* Remove all warnings and errors

**Submission Guidelines:**

| The submission should be: well commented Jupyter notebook [format - .HTML and .ipynb]  **Criteria** |  |
| --- | --- |
| **Question 1. Perform an Exploratory Data Analysis on the data**  - Univariate analysis - Bivariate analysis - Use appropriate visualizations to identify the patterns and insights - Any other exploratory deep dive |  |
| **Question 2. Illustrate the insights based on EDA**  Key meaningful observations on the relationship between variables |  |
| **Question 3. Data Pre-processing**  Prepare the data for analysis - Missing value Treatment, Outlier Detection(treat, if needed- why or why not ), Feature Engineering, Prepare data for modeling |  |
| **Question 4. Model building - Logistic Regression**  - Make a logistic regression model - Improve model performance by up and downsampling the data - Regularize above models, if required |  |
| **Question 5. Model building - Bagging and Boosting**  - Build Decision tree, random forest, bagging classifier models - Build Xgboost, AdaBoost, and gradient boosting models |  |
| **Question 6. Hyperparameter tuning using grid search**  - Tune the best 3 models using grid search and provide the reason behind choosing those models - Use pipelines in hyperparameter tuning |  |
| **Question 7. Hyperparameter tuning using random search**  - Tune the best 3 models using random search and provide the reason behind choosing those models - Use pipelines in hyperparameter tuning |  |
| **Question 8. Model Performances**  - Compare the model performance of all the models - Comment on the time taken by the grid and randomized search in optimization |  |
| **Question 9. Actionable Insights & Recommendations**  - Business recommendations and insights |  |
| **Question 10. Notebook - Overall quality**  - Structure and flow - Well commented code |  |