JJMALK Data Insights, Inc.

Model for Predicting Credit Card Fraud for BOMI

Milestone Status Report

Project Name: Model for Predicting Credit Card Fraud **Reporting Period:** October 20, 2023

Project Manager: Maricris Resma

Project Owner: JJMALK Data Insights, Inc.

Executive Summary

Overall Status: **ON TRACK**

	Status	Reason for Deviation
Budget	ON TRACK	n/a
Schedule	ON TRACK	n/a
Scope	ON TRACK	n/a

Comments:

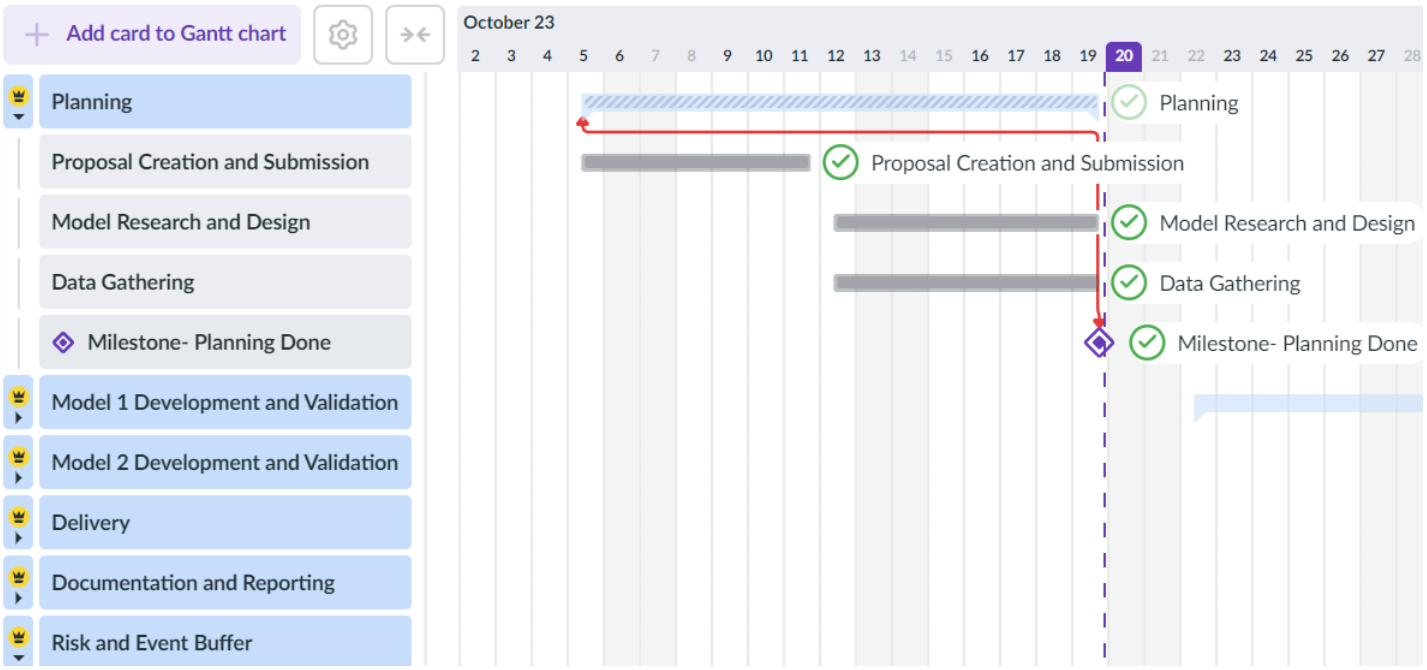
- The first Project Phase: Planning has been done
- The project **Milestone Planning – DONE** has been completed on October 20, 2023

Project Definition

Project Definition

Project Objectives	▪ Develop a predictive modelling application that effectively predicts credit card fraudulent actions
Scope	▪ Retrieve and manage credit card transaction Data from Google Cloud Platform ▪ Create a machine learning classifier model capable of detecting credit card fraud(minimum 80% accuracy, precision, recall and f1 score) using Decision Tree as the initial Model then Random Forest in case initial model does not reach desired accuracy ▪ Improve the adaptability of the model to evolving fraud patterns through continuous learning ▪ Reduce false positives to guarantee a seamless and trustworthy user experience ▪ Deliver the pickle file that can be integrated into Bank of Mississauga's organizational systems and databases
Assumptions	▪ Customer BOMI (Bank of Mississauga) will be responsible for integrating the delivered pickle file in their system

Gantt Chart of Project Timeline



Project Timeline and Deliverables

Project Tasks	Start	End	Deliverables	Status	Actual
Planning					
Proposal Creation and Submission	06-Oct-2023	12-Oct-2023		**DONE**	12-Oct-2023
Data Gathering	13-Oct-2023	20-Oct-2023		**DONE**	20-Oct-2023
Model Research and Design	13-Oct-2023	20-Oct-2023		**DONE**	20-Oct-2023
◆ ***Milestone- Planning Done***		***20-Oct-2023***	***Proposal Doc***	**DONE**	***20-Oct-2023***
Model 1 Development and Validation					
GIT Creation	23-Oct-2023	23-Oct-2023		**NOT STARTED**	
Data Preprocessing	24-Oct-2023	03-Nov-2023		**NOT STARTED**	
Exploratory Data Analysis	06-Nov-2023	10-Nov-2023		**NOT STARTED**	

Project Tasks	Start	End	Deliverables	Status	Actual
Data Modeling Using ML Algorithm 1	13-Nov-2023	22-Nov-2023		**NOT STARTED**	
Model 1 Performance Evaluation	13-Nov-2023	22-Nov-2023		**NOT STARTED**	
◆ ***Milestone - Model 1 Dev and Val Done***		***23-Nov-2023***		**NOT STARTED**	
Model 2 Development and Validation (if needed)					
Data Modeling Using ML Algorithm 2	24-Nov-2023	28-Nov-2023		**NOT STARTED**	
Model 2 Performance Evaluation	24-Nov-2023	28-Nov-2023		**NOT STARTED**	
Model Selection	29-Nov-2023	30-Nov-2023		**NOT STARTED**	
◆ ***Milestone - Model 2 Dev and Val Done***		***30-Nov-2023***		**NOT STARTED**	
Delivery					
Code Freeze	01-Dec-2023	01-Dec-2023		**NOT STARTED**	
Pickle File Generation and Delivery	04-Dec-2023	04-Dec-2023		**NOT STARTED**	
◆ ***Milestone - Customer Acceptance***		***04-Dec-2023***	***zip file of .ipynb and pickle file***	**NOT STARTED**	
Documentation and Reporting					
Documentation	05-Dec-2023	07-Dec-2023		**NOT STARTED**	
Create Presentation	05-Dec-2023	07-Dec-2023		**NOT STARTED**	
◆ ***Milestone - Presentation***		***08-Dec-2023***	***.doc report and .ppt file***	**NOT STARTED**	
Risk and Event Buffer					
Risk Buffer	11-Dec-2023	13-Dec-2023		**NOT STARTED**	
Event Buffer	14-Dec-2023	15-Dec-2023		**NOT STARTED**	

Note: ◆ *signify Key Milestones*

Accomplishments during this Reporting Period:

1. Proposal has been successfully submitted
2. Data Gathering
   - The Data was has been successfully retrieved from https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
   - There are millions of records in the source data therefore sampling will be done to take only 3 years worth of data (an estimated several thousands records) for more efficiency when training the model and testing it further
   - Description of the data is summarized in below table

Data Quality	Description
Format	.CSV
Shape	15 features 24,386,900 records Y (Dependent variable) – **Is Fraud?** X (Independent variable) -
Observations	- Imbalanced Y (29757 – Yes and 24357143 – No) - 3 features have missing values -

3. Model Research and Design
   - Machine Learning Classification Algorithm that will be used are:
     - Decision Tree
     - Random Forest
   - Following Modelling projects has been used as reference to assess best Model to use in the project
     - https://www.kaggle.com/code/imqadeerkhan/cc-fraud-detection-logistic-regression-and-ann/notebook#Splitting-of-datasets-into-test-and-training-data
     - https://www.kaggle.com/code/yichenzhang1226/ibm-credit-card-fraud-detection-eda-random-forest/notebook
     - https://www.kaggle.com/code/awaismalik1x/implementation-of-some-basic-models
   - Based on the reference below, precision is very low when using Logistic Regression and ANN therefore we will not use this for initial modelling. On the other hand, using random forest shows an precision score of 93% and recall score of 78% and F1 score of 85% which is higher than the other Models in the references above. Next, Decision Tree Model shows an precision score of .69% and recall score of .70% and F1 score of .70%.
   - Although the random forest from above references showed higher performance score overall, due to it's complexity, we have decided to use Decision Tree as the initial Model which has been our expertise in the past. Also we believe we can improve the accuracy by applying class balancing and better data preprocessing than what the other reference projects has done

4. Software, Tools and Setup
    - Programming Language that will be used to develop the Model is Python
    - IDE that will be used is Jupyter
    - For this project project, we will be using the following libraries:
        - pandas for managing the data.
        - numpy for mathematical operations.
        - sklearn for machine learning and machine-learning-pipeline related functions.
        - seaborn for visualizing the data.
        - matplotlib for additional plotting tools.

Next Steps

Plans for the next Reporting Period:

Milestone - Model 1 Dev and Val Done

1. GIT Creation
2. Data Preprocessing
    - Data profiling and cleansing will be done to ensure that missing data are handled, either by removing them or imputing values
    - Categorical features will be converted to numeric values using one-hot encoding, and numeric features will be standardized
    - Standard Scaler will be utilized to standardize the data and set them to 0 to 1 values.
3. Exploratory Data Analysis
    - Descriptive statistics will be utilized to describe the data, and charts will be used to visually summarize information
    - bivariate and univariate graphs will be used to show distribution of data
4. Data Modeling Using ML Algorithm 1
    - Decision Tree - It trains in a shorter amount of time but is still reliable in classification problems. K-fold cross-validation will be performed to determine the optimal depth of the decision tree.
5. Model 1 Performance Evaluation
    - The models will be compared based on their performance in both the training and test data.
    - The difference between the performance of the model in the training and test sets should be within 5 PPS,
    - Confusion matrix, accuracy rate, recall rate, F1 score, ROC AUC, and other metrics will be used to evaluate the models

Issue Status:

- No Issues occurred

Change Status:

- No change needed

Risk Status:

Risk	Impact Level	Mitigation Plan	Status
Imbalanced data results in a data model that may lead to bias toward the majority class and poor predictions for the minority class	High	Apply resampling to make up for the imbalanced data	***Planned***
Poor accuracy in data model due to poor quality of data may lead to difficulty that can cause delay	Medium	Apply rigorous data preprocessing to ensure removal of outliers and irrelevant data	***Planned***
The resulting performance metrics of initial modelling algorithms to be used may not reach required accuracy, causing delay in remodelling	Low	Extensive pre-study on the data and research on best model to use must be done beforehand	***Planned***
There is a chance that the project won't be completed in the allotted time due to a team member's illness	Low	The team will apply effective time management by developing efficient work breakdown where there are primary and secondary members assigned to each task	***Planned***

Notes: Status Definition

Planned *- An identified risk with a risk response plan.*

In-Process *- A risk where the risk response is being executed*

Closed *- A risk that occurred and is transfered to an issue or the risk was solved/avioded*

Not occurred *- A risk that was identified but that did not occur*

Rejected *- Created and kept for tracking pruposes but considered not to be used yet*

Budget Status as of 20/10/2023

Budget Category	Original Budget Cost	Spent to Date	Budget Left
Software Costs			
Application software	$500	$500	$0
Operating system	$1,000	$1,000	$0
Hardware Costs			
PCs	$10,000	$10,000	$0
Network Costs			
Internet	$1,000	$100	$900
Team Resources			
Project Manager ($30/hr)	$12,240	$1,200	$11,040
Technical Lead ($30/hr)	$12,240	$1,200	$11,040
Developer ($30/hr)	$12,240	$1,200	$11,040
Researcher ($30/hr)	$12,240	$1,200	$11,040
Validator ($30/hr)	$12,240	$1,200	$11,040
Total	**$ 73,700**	$17,600	$73,700

Budgetary Comments:

- Still within budget