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:** November 23, 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:

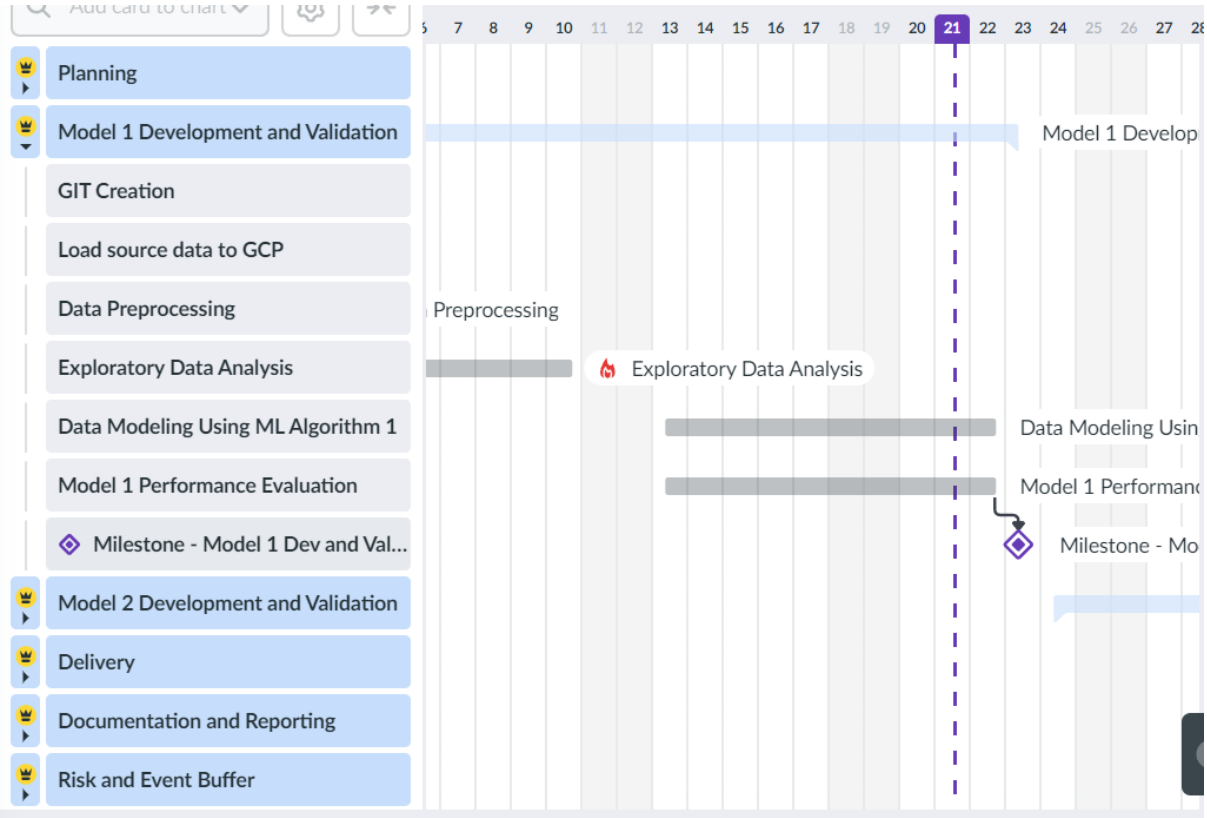
- The first Project Phase: Planning has been done
- The project **Milestone - Model 1 Dev and Val Done** has been completed on November 23, 2023

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					

Project Tasks	Start	End	Deliverables	Status	Actual
GIT Creation	23-Oct-2023	23-Oct-2023		**DONE**	23-Oct-2023
Data Preprocessing	24-Oct-2023	03-Nov-2023		**DONE**	03-Nov-2023
Exploratory Data Analysis	06-Nov-2023	10-Nov-2023		**DONE**	10-Nov-2023
Data Modeling Using ML Algorithm 1	13-Nov-2023	22-Nov-2023		**DONE**	22-Nov-2023
Model 1 Performance Evaluation	13-Nov-2023	22-Nov-2023		**DONE**	22-Nov-2023
◆ ***Milestone - Model 1 Dev and Val Done***		***23-Nov-2023***		**DONE**	***23-Nov-2023***
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. GIT Creation
2. Load data in GCP Platform
3. Data Preprocessing
   - Data profiling

Data Quality	Description
Format	.CSV
Shape	15 features 24,386,900 records Y (Dependent variable) – **Is Fraud?** X (Independent variables): # Column Dtype --- ------ ----- 0 User int64 1 Card int64 2 Year int64 3 Month int64 4 Day int64 5 Time object 6 Amount object 7 Use Chip object 8 Merchant Name int64 9 Merchant City object 10 Merchant State object 11 Zip float64 12 MCC int64 13 Errors? object
Observations	- Imbalanced Y (29757 – Yes and 24357143 – No) - 3 features have missing values ([‘Errors?’, Zip,]
Sampling	- use data in the latest year (2019). Recent information will be used to capture late st patterns in fraudulent transactions. - Retain successful transactions only (no error). - reduced to 500,000 records

   - Data cleansing

Data	Description of cleaning done
Convert features data type	Group merchant category codes
Null values	Features removed with null values
Time	Convert to float

Data Encoding	Convert categorical feature using One Hot Encoding

4. Exploratory Data Analysis
    - Histogram and Boxplot is used to show univariant data distribution of the features
    - Correaltion Heatmap showed weak multicollinearity between independent variables
    -
5. 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 performed
        i. minimum leaf size = 500
        ii. tree depth of 7 or 8 has the good combination of accuracy and recall.

6. Model 1 Performance Evaluation
    - Recall score for the 'YES Fraud' was .98 in the training data and only .80 in the test data.
    - On the other hand, Precision which is equally important shows .80 score in training data but only a mere .004 in test data which means that out of all the predictions of fraud only 4% was accurate and the rest was wrong which would cause an inconvenient experience for customer.

	1 - Yes	0 - No	weighted ave
TRAIN Model Score			
classification accuracy	NaN	NaN	0.859
precision	0.791	0.971	0.881
recall	0.978	0.741	0.859
f1 score	0.874	0.841	0.857
ROC AUC	NaN	NaN	NaN

	1 - Yes	0 - No	weighted ave
TEST Model Score			
classification accuracy	NaN	NaN	0.741
precision	0.004	1.000	0.999
recall	0.799	0.741	0.741
f1 score	0.007	0.851	0.850
ROC AUC	NaN	NaN	NaN

    - Result show overfit therefore it is recommended to use another ML model (random forest)

Next Steps

Plans for the next Reporting Period:

Model 2 Development and Validation

1. Data Modeling Using ML Algorithm Random Forest
    - K-fold cross-validation will be performed
2. Model 2 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
3. Model Selection
    - The performance from Decision Tree and Random Forest will be compared and the one that gives the best result will be chosen from where pickle file will be generated from.

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	***Closed***
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/avoided*

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	$700	$300
Team Resources			
Project Manager ($30/hr) ($240/day)	$12,240	$6,000	$6,240
Technical Lead ($30/hr)	$12,240	$6,000	$6,240
Developer ($30/hr)	$12,240	$6,000	$6,240
Researcher ($30/hr)	$12,240	$6,000	$6,240
Validator ($30/hr)	$12,240	$6,000	$6,240
Total	**$ 73,700**	$42,200	$31,300

Budgetary Comments:

- Still within budget