Bachelor of Science in Computer Science & Engineering

Machine Learning Based Fraud Incident Classification and Assistance System for Bangladeshi E-Commerce Market

by

Salman Farsi

ID: 1804102

Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)

Chattogram-4349, Bangladesh.

25th July, 2023

Chittagong University of Engineering & Technology (CUET)
Department of Computer Science & Engineering
Chattogram-4349, Bangladesh.

Thesis Proposal

Application for the Approval of B.Sc. Engineering Thesis/Project

Student Name	: Salman Farsi	Session : 2021-2022
ID	: 1804102	

Supervisor Name	: Dr. Mahfuzulhoq Chowdhury
Designation	: Associate Professor Department of Computer Science & Engineering

Department	: Computer Science & Engineering
Program	: B.Sc. Engineering

Tentative Title : **Machine Learning Based Fraud Incident Classification and Assistance System for Bangladeshi E-Commerce Market**

Table of Contents

List of Figures

List of Abbreviations

ANN Artificial Neural Network. 3, 8–10

AUC Area Under Curve. 4

CNN Convolutional Neural Network. 4

DT Decision Tree. 3, 4, 8

KNN K-Nearest Neighbor. 4, 8

LGBM Light Gradient Boosting Machine. 4, 8–10

LR Logistic Regression. 3, 4, 8

LSTM Long Short-Term Memory. 4

ML Machine learning. 3, 8, 9

RF Random Forest. 3, 4, 8–10

ROC Reciever Operating Characteristics. 4

SVM Support Vector Machine. 3, 8–10

XGBM Extended Gradient Boosting Machine. 8–10

1 Introduction

E-commerce sites are currently alternatives to brick-and-mortar shops for purchasing goods by sitting at home. During the Coronavirus Pandemic, this trend of shopping has increased more rapidly throughout the whole world. Because of the growth in e-commerce business, there are many fraudulent activities occurring in this industry. While it does happen occasionally that buyers commit fraud, it also happens more frequently when the seller commits fraud too. So, the nature of the fraudulent activities are bidirectional. Customers that commit fraud frequently do so in a variety of ways, including using stolen identities, credit cards, stealing refunds, and engaging in friendly fraud[1][2].On the other hand, the seller fraud issue is currently another major worry for the e-commerce consumer. Like the buyer, the seller has various opportunities to commit fraud. They can commit affiliate fraud, no-delivery fraud, counterfeit product fraud, triangulation fraud, and inventory fraud[2].

With a forecasted turnover of US 7,520.3 million dollars by 2023, Bangladesh acquired the 36th biggest worldwide marketplace for e-commerce, surpassing Peru[3]. The Bangladeshi e-commerce market is anticipated to develop by 13.1% in 2023, contributing to global growth at a rate of 9.6% in the same year[3]. However, amidst this booming growth, there have been some concerning incidents of online scams. E-Valy, for instance, has been involved in one of the biggest online business scams in Bangladesh, owing Tk. 500 crore[4] to its customers, according to the central bank's suggestion. Additionally, the E-orange Scam amounted to Tk. 1,100 crore[5].On the contrary, according to reports [6], small to medium-sized e-commerce platforms see a lot of client fraud concerns in Bangladesh. Customers giving fake shipping addresses when paying by cash on delivery is a common problem that bothers retailers. Furthermore, certain customers raise baseless claims about product defects, seeking unwarranted refunds, thereby causing financial losses to retailers. These occurrences underscore the urgency of implementing robust fraud detection systems from the seller's perspective, while simultaneously devising strategies to tackle customer-side fraud.

The goal of this research is to create a system based on machine learning that is capable of predicting the likelihood of fraud victimization for both customers and sellers independently. This system will classify the various types of fraud faced by individuals, such as no-delivery fraud, counterfeit product fraud, inventory fraud, and affiliate fraud on the seller's side. And identity theft, credit card fraud, and friendly fraud on the customer's side. This system will possibly be able to determine whether a fraud occurrence is genuine or not and, if confirmed, categorize the specific type of fraud to assist relevant authorities in taking appropriate and legitimate actions, with the user's consent.

Expanding the survey's participant pool to include a wider demography and maintaining its thorough coverage is one of the main difficulties of this study. Customers and sellers might not be able to immediately say if they have been the victim of fraud, therefore the dataset must be labeled while consulting fraud specialists to assure correctness and dependability. A variety of outreach tactics will be utilized to involve diverse people from varied backgrounds and areas in order to address the problem of boosting survey participation. This will make it possible to collect a thorough dataset that fairly represents the larger population.

The development of such a system might greatly minimize the instances of fraud in the e-commerce sector, promoting safer and more smooth growth for the business. The technology can boost confidence amongst buyers and sellers by successfully recognizing and preventing fraud, which will encourage more people to participate in online transactions and raise trust in e-commerce platforms as a whole leading to increased economic activity and growth.

2 Background and Present State

There are several research studies were made to detect various types of fraud in online marketplaces. Some of those are discussed in this part.

In the paper [7], the authors presented a product-oriented fraud prediction method by taking into account a number of characteristics when a consumer buys a product from an online store. They gathered the important customer, product, and seller attributes and issues through survey questionnaires given to the customers. They assessed their models using several performance criteria and discovered that certain ML models performed better in various metrics. Nevertheless, CatBoost had the best accuracy (93.28%) when there was no feature selection technique used, while SVM provided 93.09% accuracy when Extra Tree Classifier was used to choose features. However, the author in this paper overlooked some very crucial elements that could have improved the model, such as seller badge or rating, seller's offline outlet availability, product guarantee or warranty, buyer awareness of online fraud schemes, buyer caution regarding security, etc. As the customer confirms the fraud instance as their target variable when it may not have occurred at all, this research introduced bias into the dataset.

In the paper [8], a method for detecting seller or merchant fraud in an e-commerce platform was suggested. When analyzing and identifying merchant fraud on a dataset, the authors employed RF, DT, and LR algorithms based on multiple variables related to the seller and customer. By using Random Forest, they achieved the maximum accuracy of 84%. The study doesn't explain clearly how the buyer purchase behavior was also utilized to determine if the seller was fraudulent or not, and the paper also didn't explain the usefulness of some seller features as no feature extraction was done.

In the paper [9], the authors applied several Artificial Neural Network methods for the detection of online transaction fraud using various learning techniques, such as applying backpropagation, steepest descent as the first order method of learning, and quick propagation, Gauss-Newton as the second order learning method. Among all the models, the backpropagation model had the highest accuracy 96%. The limitation of this paper is that the author didn't generalize what types of fraud the seller encountered from the customer side and for the dataset containing a large number of training examples any analysis on the computational complexity of their proposed ANN was not explained through their method. Based on their feature-engineered dataset, this paper didn't investigate further ML techniques or variables that can influence fraud detection in e-commerce transactions.

In the paper [10], the authors concluded that machine learning algorithms can be effectively used to find fake e-commerce cashback transactions in Indonesia. The authors used KNN, CNN, and LSTM algorithms to detect fraud and found that the KNN algorithm performed the best with an accuracy of 83.82%. The research paper offers valuable insights into the advantages of employing machine learning classification for the detection of cashback fraud and suggests steps that can be taken to prevent fraudulent activities in the future. The study was conducted using transaction data from only one e-commerce platform in Indonesia, which may limit the generalizability of the findings to other platforms or countries. The relation between cashback fraud and other similar types of fraud like counterfeit product fraud was not addressed.

The paper [11], focused on predicting identity theft victimization and identifying key predictors of identity theft using survey data.Three machine learning algorithms, namely LR, DT, and RF, were employed for analysis. The study evaluated the efficacy of these algorithms based on criteria like overall correct classification percentage, ROC, AUC, and feature criticality.By aggregating data, the paper maximized the number of variables and cases available for analysis. The findings suggest that machine learning algorithms can effectively predict identity theft victimization and identify significant predictors of identity theft.The limitation of this study is that it does not consider the impact of demographic factors such as race and gender on identity theft victimization. The study does not examine the effectiveness of preventive measures against identity theft.

The authors of paper [12], suggested an ingenious approach for identifying credit card fraud utilizing an improved LGBM. The technique successfully identified fraudulent transactions with a high degree of accuracy, achieving the best results in terms of accuracy at 98.40%. While AUC was 92.88% and Precision was 97.34%. F1-score, however, was just 56.95%. The method used a Bayesian-based hyperparameter algorithm for optimization to tune each parameter of the machine. The efficacy of the suggested strategy for real-time fraud detection system, any investigation is not covered. Apart from the Bayesian-based hyperparameter optimization technique, exploring the use of other optimization algorithms to tune the hyperparameters of the LightGBM algorithm could increase the F1-score overall but no comparison was made on that.

The study [13] primarily investigated online shopping fraud and its focus on customer behavior and preventive actions. It relied on secondary survey data to analyze how individuals protect themselves from online shopping fraud. The research underlined the significance of enhancing prevention advice for online shoppers to mitigate the risk of fraud. The findings indicated that many respondents do not implement appropriate measures to reduce their risk during online shopping, underscoring the need for improved online safety guidance. However, one limitation of the study is the lack of insights from individuals who have experienced online shopping fraud, which could have provided valuable input for devising effective prevention strategies.

In conclusion, It can be said that for classifying a specific form of fraudulent incident out of all, no study has been done yet on the privilege of fraud prevention. Besides no research was done which worked on both the customer and seller by incorporating several of their attributes and transaction behavior to predict whether they actually became the victim of a fraudulent incident or not. Hence this study will focus on these issues that are missing and making an overall prediction system that is helpful for both seller and customer in taking action based on the type of fraud victim they are.

3 Specific Objectives and Possible Outcomes

The main objectives of this work are as follows:

1. To collect datasets from customers and sellers using survey-based questionnaires, by ensuring a wider demographic coverage of fraudulent incidents.

2. To develop a machine learning model capable of predicting the likelihood of e-commerce users falling victim to fraud and classifying the specific type of fraudulent case each individual has encountered.

3. To develop an Android application that serves as a unified platform for customers and sellers to communicate and settle unsatisfactory transactions

effectively.

4. To enable users to take legal action against fraud incidents, providing facilities for both sellers and customers to report fraud types predicted by the proposed model to the appropriate legal authorities, supported by essential evidence.

4 Outline of Methodology

The primary aim of this study is to develop a model that can accurately classify the fraud incident type for both seller and customer separately based on a range of customer and seller features. To illustrate the functionality of this proposed model, a visual representation is provided in figure 4.1. This model holds the potential to significantly enhance fraud verification and classification, leading to a more secure and trustworthy e-commerce environment.

4.1 Data Collection

For the dataset, survey-based questionnaires will be taken from both the customer and seller independently focusing on the general demographic information, fraud awareness, and knowledge, product features, several seller's or customer's characteristics, payment method, and possible fraud occurrence-related information [14] [15]. Here two datasets will be collected through several questionnaires on the customer and seller separately. In the customer dataset, the possible features of the dataset will be age, gender, occupation, education level, online fraud scheme knowledge, date of purchase, purchased product type, product discount, carefulness about product specification, product price, acknowledging seller review and rating, seller's shop age, payment method, the problem of the product, what fraud occurred, evidence of defective product, etc. On the other hand, in the seller dataset, the possible features of the dataset will be the seller's shop age, average revenue per year, knowledge of fraud scheme, fraud scheme in action, buyer information check, fraud incident-related information, and other relevant features regarding the fraudulent scenario.

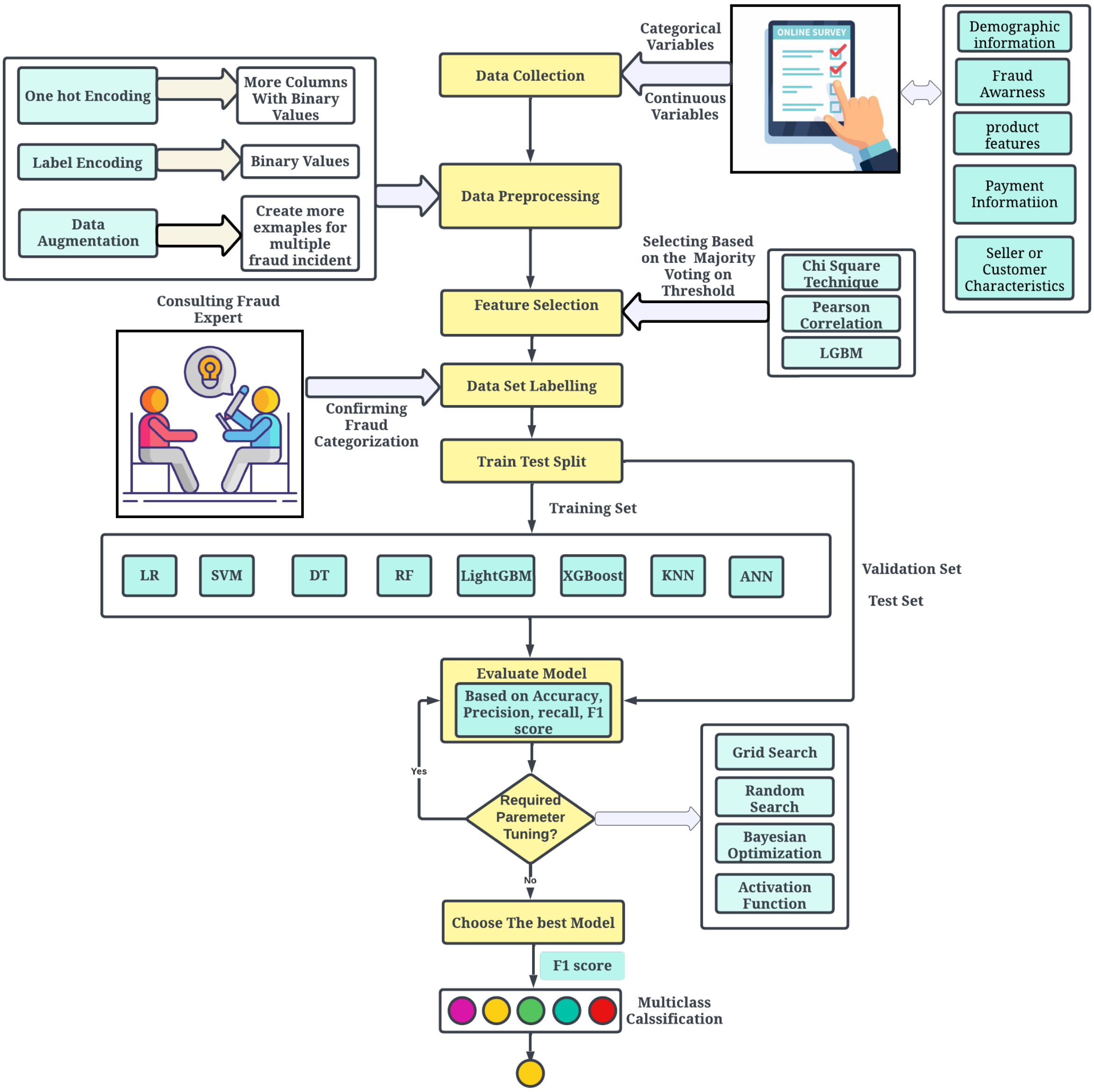


Figure 4.1: Process flow diagram of the proposed methodology

4.2 Data Preprocessing

Since the data was gathered through a survey, preprocessing is necessary beforehand. In order to preprocess the data, it should be first checked to see whether any record has an empty cell or not. For the sake of this situation, it will be supposed that the answer to that question is not applicable to that respondent when the response is categorical. But for continuous values like the price of the product, discounted percentage, or cashback amount offered, the mean value will be taken in place of the empty cell. Finally, to convert the categorical variables into numerical values, One Hot Encoding [10] will be used. On the contrary, for binary variables, Label Encoding will be used.

4.3 Feature Selection

To analyze the substantially contributing features, three algorithms will be used to pick the features. Pearson Correlation, Chi-Square Technique [10], and LightGBM [12] are the feature selection techniques that will be employed. After every feature that passes the three algorithms' test, will be collected and a vote will be made to decide which features could be used in this task to train the model.

4.4 Dataset Labeling

Each data point will be labeled by assigning the appropriate fraud category based on the user responses. To label the dataset, consultation with fraud experts is needed. This will be the target variable or class labels for the proposed machine learning model.

4.5 Train-Test-Split

The dataset will be split into three sets. The first one is training, the second one is testing, and the final one is validation set. The machine learning model will be trained using the training set. The constructed model will then be assessed on the set used for validation to determine how well it performs after being trained on the set for training. Based on the performance of the validation set, the model's hyperparameters[16], which govern how it behaves during training, may be modified to enhance generalization. The performance of the testing set will then be assessed. The test set will be utilized to assess the ML model's ultimate performance following all training and fine-tuning procedures.

4.6 Training the Model

For this research work, several machine learning models such as LR, SVM,RF, DT, XGBM, LGBM, KNN and ANN are chosen initially. As seen by several previous works, these are performing better in predicting both binary and multiple classes [7] [9] [12]. The reason for using tree ensemble methods DT, RF, XGBM and LGBM is that in paper [7] the authors found better results with tree ensemble method for questionnaires dataset which heavily contained binary values in the feature. On the other hand, in one of the fraud detection paper [9] ANN outperform others. In the paper [10], the authors got higher accuracy by KNN in predicting cashback fraud transactions. As the study suggests [16], by fine-tuning the hyperparameters of the decision tree algorithm, such as adjusting the tree's maximum depth, minimum samples per leaf, and splitting criteria, the performance on a dataset primarily composed of binary values (0 or 1) can be enhanced. Random Forest can effectively handle class imbalances since each tree is trained on a different subset of data, making it more likely to capture patterns in minority classes. So, in the case of an imbalanced dataset, the uses of RF will probably give a more accurate classification [16]. In the paper [17], by applying the different kernel tricks and parameter tuning, the authors effectively used SVM to get higher accuracy. All of these scenarios suggested that the use of the aforementioned ML model for this work may provide better performance and the necessary changes in the model by parameter tuning will give a performance boost.

4.7 Model Evaluation

For the model evaluation, all the ML models that will be used in this study can be compared by several scores like accuracy, recall, precision, and F1 score. After the comparison, the best model will be chosen for final prediction and further deployment in the Android application.

4.8 Parameter Tuning of the Model

If the models' performance is not satisfactory then the hyperparameter tuning [7] will be done by changing several parameters of each model to get the best accuracy from the model. Also by observing the bias and variance issue, necessary steps will be taken to improve the model until a desired result is achieved. Grid search may be used to create a grid of possible hyperparameter values, and the procedure of tuning assesses how well the model performs for each set of possible hyperparameters. Besides Random Search and Bayesian optimization can also be done to optimize models like SVM, RF, XGBM, LGBM, and ANN. Changing several activation values in the ANN along with the addition of several layers is another way that will be used in the experiment of getting higher performance.

4.9 Choosing the Best Model

In order to choose the best model, the F1 score will be considered because it confirms that the model is classifying the classes well. In particular scenarios, where there is an imbalance between the classes, a higher F1 score will provide greater details about classified and miss-classified classes.

4.10 Final Classification

The best model will be used for the final prediction and for taking the user input to predict the fraudulent class. The class will be different for the buyer and seller since the models will be run individually for both buyer and seller. The understanding of these types of fraud came through background reading [1] [2] [12] [7].

4.10.1 Seller Sided Fraud Class

- No Fraud Occurred
- Affiliate fraud
- No-delivery fraud
- Counterfeit product fraud
- Triangulation fraud
- Inventory fraud

4.10.2 Customer Sided Fraud Class

- No Fraud Occurred
- Credit card fraud
- Identity theft

- Account Takeover Fraud
- Price Manipulation
- Friendly Fraud

5 Android Application

An Android application will be developed to enable the user to see what type of fraudulent incident they have encountered and to check the validity of their allegation. There are several other features that will eventually be included in the application so that it becomes more user-friendly and effective. The process flow diagram of the proposed Android application is shown in figure 5.2 and 5.3.

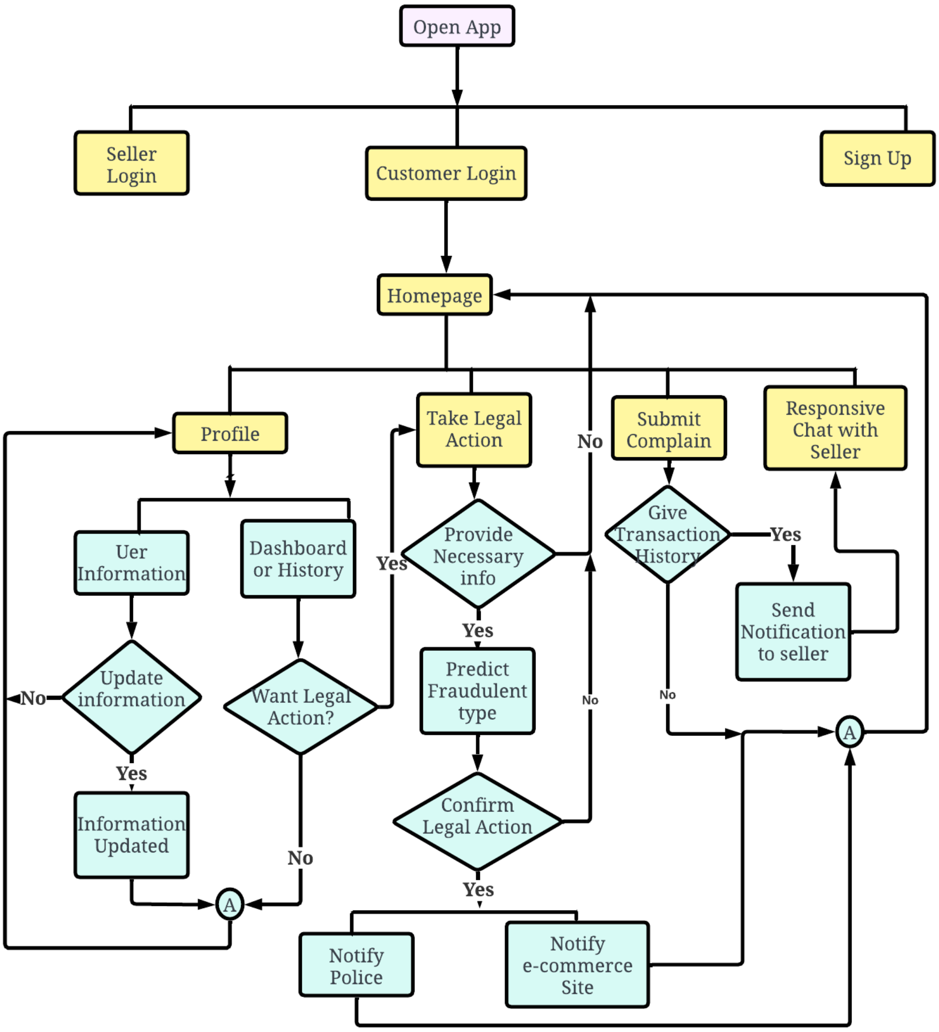


Figure 5.2: Process flow diagram of the customer interface

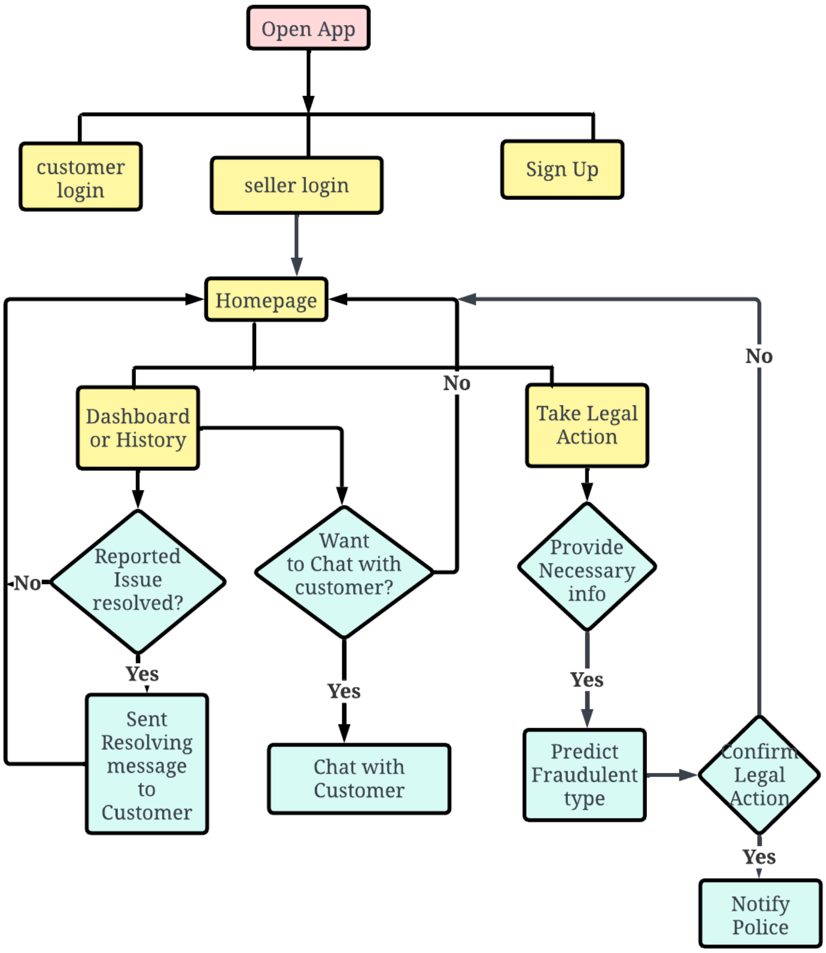


Figure 5.3: Process flow diagram of the seller interface

- The app will allow the user to complain about their issues to the registered seller before taking legal action.
- The user can directly chat with the seller after the complaint is submitted so that they can clarify the issue among themselves.
- If the seller doesn't hear to the customer's allegation by providing necessary details and answering questionnaires they will be able to see what type of fraud has occurred with them.
- Based on the prediction they can directly submit legal action to police or relevant authority. In the dashboard, the customer can see his action history and status.
- As soon as the action is taken, a notification will be sent to the seller. And a complaint ID will be assigned.

- But if the issue is resolved, then both the customer and seller will confirm it, and in the dashboard of both the customer and seller it will be shown as a resolved issue.

6 Required Resources

6.1 Hardware Resources

- Personal Computer
- Android Device

6.2 Software Resources

- OS: Windows 10
- Tools: Anaconda, Visual Studio, Flutter, Firebase, Google Map API
- Libraries: Tensorflow, Keras, Numpy, Pandas, Matplotlib, Sklearn
- Programming Language: Python, Drat

7 Cost Estimation

The cost estimation of this research work is given below. This cost estimation may need to be changed to the requirements.
a. Cost of Data Collection :

• Internet Cost	Tk 2000
• Survey Cost	Tk 50000
Total	Tk. 52000

b. Cost of Materials :

• Hardware Equipments	Tk 100000
• Softwares	Tk 25000

Total	Tk. 125000

c. Drafting & Binding :

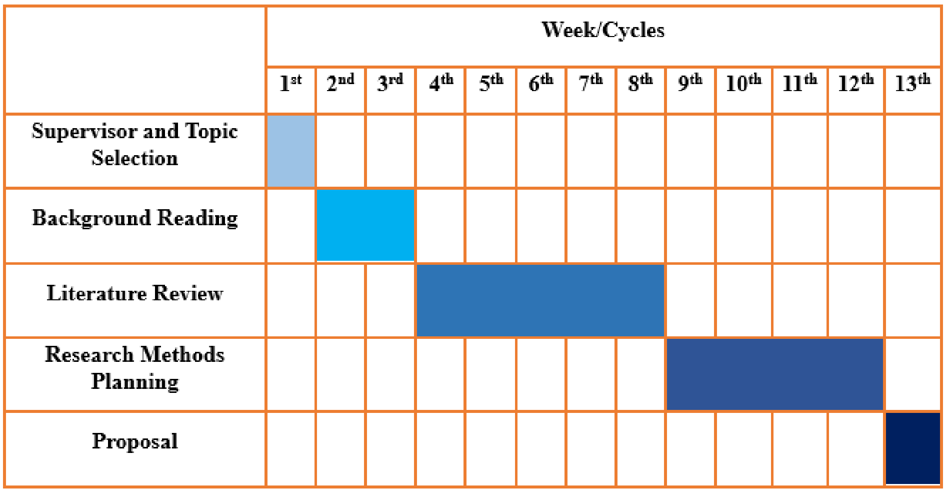
• Paper	Tk 500
• Drafting	Tk 1000
• Printing	Tk 1000
• Binding	Tk 500

Total	Tk. 3000
Miscellaneous	Tk. 1000

Grand Total	Tk. 181000

7.1 Time Management

Gantt Chart for the entire timeline from the beginning of the thesis to the end is added here.



	Week/Cycles												
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	12th	13th
Supervisor and Topic Selection	■												
Background Reading		■	■										
Literature Review				■	■	■	■	■					
Research Methods Planning									■	■	■	■	
Proposal													■

Figure 7.4: Gantt chart of time management.

CSE Undergraduate Studies (CUGS) Committee

Reference :

Meeting No : **Resolution No :** **Date :**

____

Signature of the Student

____

Signature of the Supervisor

____

Signature of the Head of the Department

References

[1] Richard, *E-COMMERCE FRAUDS – COMMON TYPES AND PREVENTION TIPS*, `https://shuftipro.com/blog/e-commerce-frauds-common-types-and-prevention-tips/`, [Online; visited on 25th May 2023], Sep. 2020 (cit. on pp. 1, 10).

[2] Francesca, *The Complete Guide to Ecommerce Fraud Prevention*, `https://ecommerceguide.com/guides/ecommerce-fraud/`, [Online; visited on 25th May 2023], May 2017 (cit. on pp. 1, 10).

[3] *eCommerce market in Bangladesh*, `https://ecommercedb.com/markets/bd/all`, [Online; visited on 5th July 2023], 2023 (cit. on p. 1).

[4] *The rise and fall of Evaly*, `https://www.tbsnews.net/economy/rise-and-fall-evaly-303613`, [Online; visited on 7th June 2023], Sep. 2021 (cit. on p. 1).

[5] S. Islam, *TK 1,100cr Eorange Scam: Blame-shifting fails*, `https://www.thedailystar.net/business/economy/e-commerce/news/tk-1100cr-eorange-scam-blame-shifting-fails-2189996`, [Online; visited on 7th June 2023], Oct. 2021 (cit. on p. 1).

[6] Q. T. Islam and N. I. Saeed, *E-commerce in Bangladesh: prospects and challenges*, `https://www.newagebd.net/article/147734/e-commerce-in-bangladesh-prospects-and-challenges`, [Online; visited on 7th June 2023], Aug. 2021 (cit. on p. 1).

[7] M. Sabih *et al.*, 'Fraud prediction in pakistani e-commerce market,' in *2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT)*, 2021, pp. 01–06. DOI: `10.1109/ISAECT53699.2021.9668438` (cit. on pp. 2, 8–10).

[8] F. Hasan, S. K. Mondal, M. R. Kabir, M. A. Al Mamun, N. S. Rahman and M. S. Hossen, 'E-commerce merchant fraud detection using machine learning approach,' in *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, 2022, pp. 1123–1127. DOI: `10.1109/ICCES54183.2022.9835868` (cit. on p. 3).

[9] S. Alqethami, B. Almutanni and M. AlGhamdi, 'Fraud detection in e-commerce,' *International Journal of Computer Science & Network Security*, vol. 21, no. 6, pp. 312–318, 2021 (cit. on pp. 3, 8, 9).

[10] B. Karunachandra, N. Putera, S. R. Wijaya, D. Suryani, J. Wesley and Y. Purnama, ‘On the benefits of machine learning classification in cashback fraud detection,’ *Procedia Computer Science*, vol. 216, pp. 364–369, 2023, 7th International Conference on Computer Science and Computational Intelligence 2022, ISSN: 1877-0509. DOI: `https://doi.org/10.1016/j.procs.2022.12.147` (cit. on pp. 3, 7–9).

[11] X. Hu, X. Zhang and N. Lovrich, ‘Forecasting identity theft victims: Analyzing characteristics and preventive actions through machine learning approaches,’ *Victims Offenders*, vol. 16, pp. 465–494, Apr. 2021. DOI: `10.1080/15564886.2020.1806161` (cit. on p. 4).

[12] A. A. Taha and S. J. Malebary, ‘An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine,’ *IEEE Access*, vol. 8, pp. 25 579–25 587, 2020. DOI: `10.1109/ACCESS.2020.2971354` (cit. on pp. 4, 8, 10).

[13] J. Whittaker, M. Edwards, C. Cross and M. Button, ‘"i have only checked after the event": Consumer approaches to safe online shopping,’ *Victims and Offenders*, pp. 1–23, Oct. 2022. DOI: `10.1080/15564886.2022.2130486` (cit. on p. 5).

[14] C. Cross, M. Edwards, M. Button and J. Whittaker, *Online shopping fraud victimisation in Australia*, `https://www.researchgate.net/publication/366167476_Online_shopping_fraud_victimisation_in_Australia`, [Online; visited on 25th May 2023], Oct. 2022 (cit. on p. 6).

[15] M. Zeng, H. Cao, M. Chen and Y. Li, ‘User behaviour modeling, recommendations, and purchase prediction during shopping festivals,’ *Electronic Markets*, vol. 29, Sep. 2018. DOI: `10.1007/s12525-018-0311-8` (cit. on p. 6).

[16] I. Sarker, ‘Machine learning: Algorithms, real-world applications and research directions,’ *SN Computer Science*, vol. 2, Mar. 2021. DOI: `10.1007/s42979-021-00592-x` (cit. on pp. 8, 9).

[17] V. Blanco, A. Japón and J. Puerto, ‘Optimal arrangements of hyperplanes for svm-based multiclass classification,’ *Advances in Data Analysis and Classification*, vol. 14, Jul. 2019. DOI: `10.1007/s11634-019-00367-6` (cit. on p. 9).