Sri Lanka Institute of Information Technology

Fundamentals Of Data Mining (IT3051)

Continuous Assignment – 2023, Semester 2

SOW

A picture containing text, clipart, vector graphics

Description automatically generated

Zainab M.Z IT21070594

Rashida M.S.F IT21013850

Nuha M.N IT21004568

Bishirhafi F.S.M.T IT21006166

**Problem Statement: Improving Credit Card Fraud Detection for Enhanced Financial Security**

**Background:**

In the modern digital age, credit card transactions play a pivotal role in everyday financial activities. However, with the increasing prevalence of cybercrime, credit card fraud has become a significant concern for both cardholders and credit card companies. To safeguard customers from unauthorized charges and protect financial institutions from losses, it is imperative to enhance the accuracy and efficiency of credit card fraud detection systems.

**The Challenge:**

The challenge at hand is two-fold. On one hand, credit cardholders face the risk of unauthorized charges, potentially leading to financial losses and personal inconveniences. On the other hand, financial institutions, including credit card companies and banks, must contend with substantial financial losses resulting from fraudulent transactions. The increasing sophistication of fraudsters and the diversity of fraudulent activities make it clear that traditional methods of fraud detection are no longer sufficient to protect the interests of both cardholders and financial institutions.

**The Importance of Accurate Fraud Detection:**

Enhancing the accuracy and efficiency of credit card fraud detection systems is imperative. Timely identification of fraudulent transactions is not only crucial for preventing unauthorized charges but also for preserving the trust and confidence of cardholders in the financial system. Moreover, for financial institutions, minimizing losses due to fraud is essential to maintain profitability and competitiveness.

**Dataset Description:**

The dataset provided contains credit card transactions made by European cardholders in September 2013. Among the 284,807 transactions recorded over a two-day period, there were 492 cases of fraudulent transactions. The dataset is highly imbalanced, with fraudulent transactions accounting for only 0.172% of the total.

[Data Set](https://drive.google.com/drive/folders/1MZWEmRZFHMvglwNqYTrl797eb3holptB)

**Scope of Work**

We will analyze this data to:

* Identifying fraudulent transactions- The main objective is to detect transactions that are not legitimate, potentially involving stolen credit card information or other forms of fraudulent activity.
* Reducing false positives- While detecting fraud is crucial, it's also important to minimize false positives, which are legitimate transactions incorrectly flagged as fraudulent. This is because blocking legitimate transactions can lead to customer dissatisfaction.
* Real-Time Detection: Many systems aim to identify fraud in real-time, preventing fraudulent transactions from being approved or allowing them to be approved with additional verification.
* Pattern Recognition: Analyzing historical transaction data can help identify patterns or anomalies associated with fraudulent activity. Machine learning algorithms can be used to detect these patterns and adapt over time as new fraud techniques emerge.
* The different algorithms that we will be using to build classification models are as follows:

1. Decision Trees
2. Support Vector Machines
3. Neural Networks
4. Random Forest
5. Naïve Bayes
6. KNN (K-Nearest Neighbors)

* As the final output of the project, a Web Application will be delivered to the end user, where the end user can view the results of the dataset based on different models/techniques used.

**Activities**

1. Select a domain, define the problem statement, and identify datasets relevant to the problem.
2. Segregate the problem as per its nature and select appropriate data mining techniques.
3. Select suitable tools and technologies to perform the required tasks.
4. Build the classification models based on the algorithm selected.
5. Compare and assess two algorithms and select the best with the accuracy and performance using a common metric, such as accuracy.
6. Design user interfaces for the web application and develop final application along with the data mining models.
7. Deploy the application and perform testing based on test cases defined.
8. Create a report consisting of all the key information on the project, its process and final output.
9. Create a video to demonstrate the application to users and interested stakeholders.

**Approach**

|  |  |
| --- | --- |
| Purpose | Tools/Methodology |
| Datasets | From Kaggle |
| Data Mining Technique and Algorithms | Classification: -  Decision Trees  Support Vector Machines  Neural Networks  Random Forest  Naïve Bayes  KNN (K-Nearest Neighbors) |
| Programing Language | Python |

**Note : Technologies and algorithms are subject to change.**

**Deliverables**

* 1. Statement of Work (SOW) Report- The report consists of key information on the project background, problem definition, scope of work, tools and techniques used, activities planned along with timeline and deliverables, also the tasks performed by each member.
  2. Data Mining Models - The model that would perform the tasks (predictive or segmentation) for the selected data sets. This may be a file consisting only the code relevant to the models developed.
  3. Web Application - Hosted web application that could be assessed by the users to perform the operations. The model(s) are integrated to the web application.
  4. Video - Demonstration of the features of the application.
  5. Final Report- The report consists of information on project domain, solution developed and test cases.

**Project Plan and Timeline**

* Find what are the latest real-world problems which should be addressed through a Data Mining task. [08-09-2023]- [09-09-2023]
* Find a latest dataset [10-09-2023]
* Create and submit the statement of work document and getting the approval for the project [22-09-2023] – [24-09-2023]
* Data Cleaning, Preprocessing, and all the Data Engineering tasks [24-09-2023] – [26-09-2023]
* Building the Models [27-09-2023] – [03-10-2023]
* Evaluate the models [03-10-2023] – [06-10-2023]
* UI/UX Designing and Deploying the models [06-10-2023] – [10-10-2023]
* Creating the Report [11-10-2023] – [14-10-2023]
* Submitting the project [15-10-2023]

**Assumptions**

• The columns that were removed do not affect the prediction of the final output.

• The audience of the Final Product is someone with the proper knowledge to understand and get a clear idea of the system outcome.

• The users have the values for the required inputs for the system to predict the values.

• The dataset is being updated annually as mentioned in the source.

**Project Team, Roles and Responsibilities**

|  |  |  |
| --- | --- | --- |
| **Member** | **Role** | **Responsibilities** |
| Zainab M.Z | Data Scientist  Data Engineer  UI/UX Engineer  Data Analyst | * Data cleaning and preprocessing. * Building the models. * Testing and Evaluating. * Design UI/UX and integrating the project * Visualize the results * Creating the final report |
| Rashida M.S.F | Data Scientist  Data Engineer  UI/UX Engineer  Data Analyst | * Data cleaning and preprocessing. * Building the models. * Testing and Evaluating. * Design UI/UX and integrating the project * Visualize the results * Creating the final report. |
| Nuha M.N | Data Scientist  Data Engineer  UI/UX Engineer  Data Analyst | * Data cleaning and preprocessing. * Building the models. * Testing and Evaluating. * Design UI/UX and integrating the project * Visualize the results * Creating the final report |
| Bishirhafi F.S.M.T | Data Scientist  Data Engineer  UI/UX Engineer  Data Analyst | * Data cleaning and preprocessing. * Building the models. * Testing and Evaluating. * Design UI/UX and integrating the project * Visualize the results * Creating the final report |