Task 2 :Business understanding

Identifying business goals

Background:

Credit card fraud refers to unauthorised use of someone's credit card information to make transactions on their behalf. Nowadays credit card fraud can happen in many different ways. Not only stolen physical cards are used for fraudulent transactions, but also online hacking or or phishing attacks are used, to gain access to credit card information of a victim. With hijacked credit card details, the illegal transaction can happen even months after the data was obtained by the criminals, making it more difficult for the victim to discover that their bank account has been breached. According to the Estonian Police and Border Guard service, the amount of financial damage caused by different kinds of online bank, card and credit card theft reached about 2 million euros in the year 2022[1].

1) https://www.politsei.ee/et/uudised/kelmuse-ohvriks-langemisest-paeaestab-inimese-enda-nutikus-10999

Business goals:

The goal of the given project is to use data science methods on a dataset of transaction data, to identify unusual transaction patterns for a client and try to label the transaction as fraudulent or not. For that a machine learning model will be trained. During the process the efficiency of different models will be tested with a range of data from the dataset. The complexity of the model will also be considered, the goal is to use as few data parameters as possible to make accurate predictions, to lessen the load of training the model with data that could have minimal impact on the final outcome.

Business success criteria:

The project would be considered successful if the trained machine learning model could label a transaction as fraudulent with accuracy of $x>66\%$. With that the majority of fraudulent transactions could be processed and rejected by the bank in time so that no greater damage could be done to the victim.

Assessing the situation

Inventory of resources:

For the given project a synthetically generated dataset of credit card transaction data will be used. The data was generated using Sparkov Data Generation tool created by Brandon Harris [2]. The project team consists of three people, who will use their laptop computers for exploring the data and training the model. Jupyter notebook will be used as the development environment.

2) https://www.kaggle.com/datasets/kartik2112/fraud-detection?select=fraudTrain.csv

Requirements, assumptions and constraints:

The planned time of completion of the project is on 10.12.2023. As the data used in the project is synthetic, there are no additional security obligations to be considered.

Risks and contingencies:

For the group project, various working environments can be considered, given technical difficulties are presented. Other causes for delays may include the vast amount of homework and assessments the project members have to attend to during the end of the semester, which can reduce the quality of the outcome.

Terminology:

There is no specialty specific terminology in the dataset that currently needs to be explained.

Costs and benefits:

There are no monetary costs needed for the completion of this project. The benefit of the project is the trained machine learning model that could predict fraudulent transactions. In addition the knowledge discovered from the data could also be considered beneficial.

Defining the data-mining goals

Data-mining goals:

The goal of the given project is to produce a predictive model that could label credit card transactions as fraudulent or legitimate based on the input data.

Data-mining success criteria:

The project would be considered successful if a predictive model with an accuracy of greater than 66% would be produced.

Task 3. Data understanding (1 points)

Data requirements

For our project we would need data in preferably .csv format, where every row is a separate transaction and where every transaction is classified as fraudulent or not. In an ideal world, we would like data from a time range as wide as possible to also observe trends over time, but considering that getting access to real credit card transaction data is very difficult, we are using synthetic data. That means that the time range does not play as big a role in our project as all data is randomized and unaffected by real world events.

Data availability

All data we are using is accessible to us through Kaggle, where it is available for free. This is due to the aforementioned high confidentiality of real credit card transaction data, which we do not have access to at the moment.

Selection criteria

There is a wide range of datasets about credit card fraud available on Kaggle. Our choice leans towards a large synthetic dataset so that we can better analyse the impact of each value on the outcome. In the cases of real world data that we found, such as Credit Card Fraud Detection Dataset 2023 by Nidula Elgiriyewithana, all data was already anonymised with nondescript column names and normalised data. We picked Credit Card Transactions Fraud Detection Dataset by Kartik Shenoy with more than a million rows of data in the train dataset to ensure we still have a large number of entries even in the event we have to weed some rows out or do some oversampling to better train our model. We are most interested in columns that are to do with the location, time and merchant information, and less interested in the personal details, such as names, of the card holders, as someone's last name is generally not correlated with their personality or behaviour.

Data description

In this dataset, the data is already separated into training and testing datasets for us. There are 22 columns in the dataset: 'trans_date_trans_time' – the time and date in a human readable format, 'cc_num' – the number on the used credit card, 'merchant' – the merchant where the transaction was committed, 'category' – the category of business for the aforementioned merchant, 'amt' – the used amount of money, 'first' – the first name of the cardholder, 'last' – the last name of the cardholder, 'gender' – the gender of the cardholder, 'street' – the street address of where the transaction was committed, 'city' - the city of the transaction, 'state' – the state of the transaction, where US state name abbreviations are used,, 'zip' – the zip code for the area of the transaction, 'lat' – the latitude coordinate of the transaction location, 'long' – the longitude of the transaction location, 'city_pop' – the population of the city where the transaction was committed, 'job' – the job of the cardholder, 'dob' – the cardholder's date of birth, 'trans_num' – the assigned number of the transaction, 'unix_time' – the time of the transaction in unix time for machine readability, 'merch_lat' – the latitude coordinate of the merchant's location, 'merch_long' – the longitude coordinate of the merchant's location, and, of course, 'is_fraud' – shows whether that transaction was found to be fraudulent or not. All fields are numerical other than 'merchant', 'category', 'gender', 'dob', 'first' and 'last' pertaining to names, 'street', 'trans_num', which is an alphanumeric code in string form, and 'job'. 'Category' only has 14 different labels across more than a million rows, so it is easy to label encode it for future work.

Data exploration

We used the describe() method for pandas dataframes to explore our data. There seem to be no initial problems with the data, especially for the numerical fields. However, the range in values for most columns seems to be rather high, which is to be expected for very specific values such as geographic coordinates or credit card numbers. As this is a synthetic dataset, there are no cases on unknown values in this dataset, which lessens the workload on data cleanup. Still, it is obvious normalising is necessary in this case to bring all values closer together for better model training later on.

Verification of data quality

From our data exploration we have concluded that this dataset is indeed capable of satisfying our requirements for this project. There are no glaring issues with the dataset and it will be ready for use in our project after minor cleanup and normalisation.

Project planning

- Normalising and cleaning of data, removing unnecessary attributes, improving usability for analysis and usage for machine learning models (5h Maare)
- Data visualisation - what trends can we see from the data straight away.(7h Karoliina)
- Preparing data for machine learning models- oversampling, undersampling, PCA, one hot encoding (3h Enriko)
- Training and validation of different machine learning models, hyperparameter tuning (ca 65h Maare, Karoliina, Enriko)
    - Training a RandomForest classifier seems to be one of the most straightforward approaches as the label we want to predict is a binary value
    - Also exploring the K-Neighbours clustering with PCA could be explored
    - SVM
    - With the most promising models hyperparameter tuning will be performed
- Model validation
- Testing the model on the test dataset.(2h Maare)
- Poster creation for presenting the results (5h Enriko)

As the training data set is fairly large, the most time will be spent on model training and hyperparameter tuning. Also the given plan is a rough estimate and team members will also help and fill in by tasks that are not their direct responsibility. Teamwork makes the dream work.