**Project Proposal - Creating Value Through Data Mining**

*Project authors: Purity Gikonyo, Artem Demidov, Hannes Guth*

**Overview**

This project is about predicting online payment fraud as a dependent variable from other measures of the given dataset. To achieve this, the dataset will be explored and analysed in the first step. Due its size, we will explore possible ways to sample the dataset or introduce ways how to deal with the number of observations. Then different machine learning algorithms will be applied, first supervised learning and if applicable and useful

(what will be found out during the analysis) also unsupervised.

We will give a reflection of the used methods and their utility in order to achieve accurate predictions in terms of previously set baseline expectations/goals.

**Goals**

* Figure out a possible relationship between several characteristic of an online payment and the payment being fraud
* Build machine learning models to predict this
* Answer questions like
  + Are there origins/destinations that are especially prone to fraud?
  + Are the previously set results/expectations met by our analysis?

**Data Overview** (description, pedigree, dimensions) and link to the data source

* The dataset contains transactions of payments made online, described by 10 features. The type of payments were either in form of cash out, transfer, debit or payment. This dataset presents transactions, where we have 8,213 frauds out of 6,362,620 transactions. The dataset is highly unbalanced, the positive class (frauds) accounts for 0.129% of all transactions.
* The source of the dataset is *kaggle*. It was issued initially on medium.com.
* Neither on kaggle nor *medium.com* there is information given about the other aspects of the pedigree.
* Link to the dataset: *https://www.kaggle.com/datasets/jainilcoder/online-payment-fraud-detection?select=onlinefraud.csv*

**What is the outcome variable and what is the task to perform: classification or regression (prediction) - for supervised learning**

The outcome variable is the *isFraud* value. It is binary. Therefore, the task to perform is a classification.

**Steps of the analysis**

* Understanding - Project objectives and Requirement Understanding
* Exploring & Understanding Data – initial data collection & familiarisation
* Data Preparation – data cleaning, exploratory data analysis, data transformation
* Modeling - selecting, building and training the models (as listed below) and predict the dependent variable
* Model Evaluation – review results, find the best fit model(s)
* Deployment – Deploying the best fit model(s) to derive predictions and insights.

**Machine Learning Methods**

* Knn
* Logistic regression
* Neural nets
* Classification tree
* Ensembles
* Naive bayes

**Structure of the final report**

* Introduction (aim, dataset description (qualitative), baselines, content introduction etc.)
* Data preparation (import, sample, clean, NAs, dimension reduction, standardisation, possibly add/calculate extra features)
* Exploratory analysis (autocorrelations, visualisations, review of baselines, dealing with unbalanced classes, add statistics about the data)
* Machine Learning part
  + Baseline models
  + Apply methods from above
  + If applicable and useful: Unsupervised learning part
* Conclusion
  + Technical conclusion: Which methods performed well, baselines kept?, ...
  + Content conclusion: Which features are useful to predict fraud, what could be learned in general, e.g. suggestions how to prevent fraud

**Specifications**

Dataset consists of 6.4 million transactions, each transaction is described by 10 features.

In the following table, “Feature” and “Description” were directly taken from kaggle.

| **Feature** | **Description** | **Type** |
| --- | --- | --- |
| step | represents a unit of time where 1 step equals 1 hour | int |
| type | type of online transaction | string |
| amount | the amount of the transaction | float |
| nameOrig | customer starting the transaction | string |
| oldbalanceOrg | balance before the transaction | float |
| newbalanceOrig | balance after the transaction | float |
| nameDest | recipient of the transaction | string |
| oldbalanceDest | initial balance of recipient before the transaction | float |
| newbalanceDest | the new balance of recipient after the transaction | float |
| isFraud | fraud transaction | bool |