**Credit Card Fraud Prediction**

Milestone 2: Data Selection and Project Proposal

Vinay Nagaraj & Vikas Ranjan

DSC630, Spring 2021

Bellevue University, NE

**Introduction**

**Background**

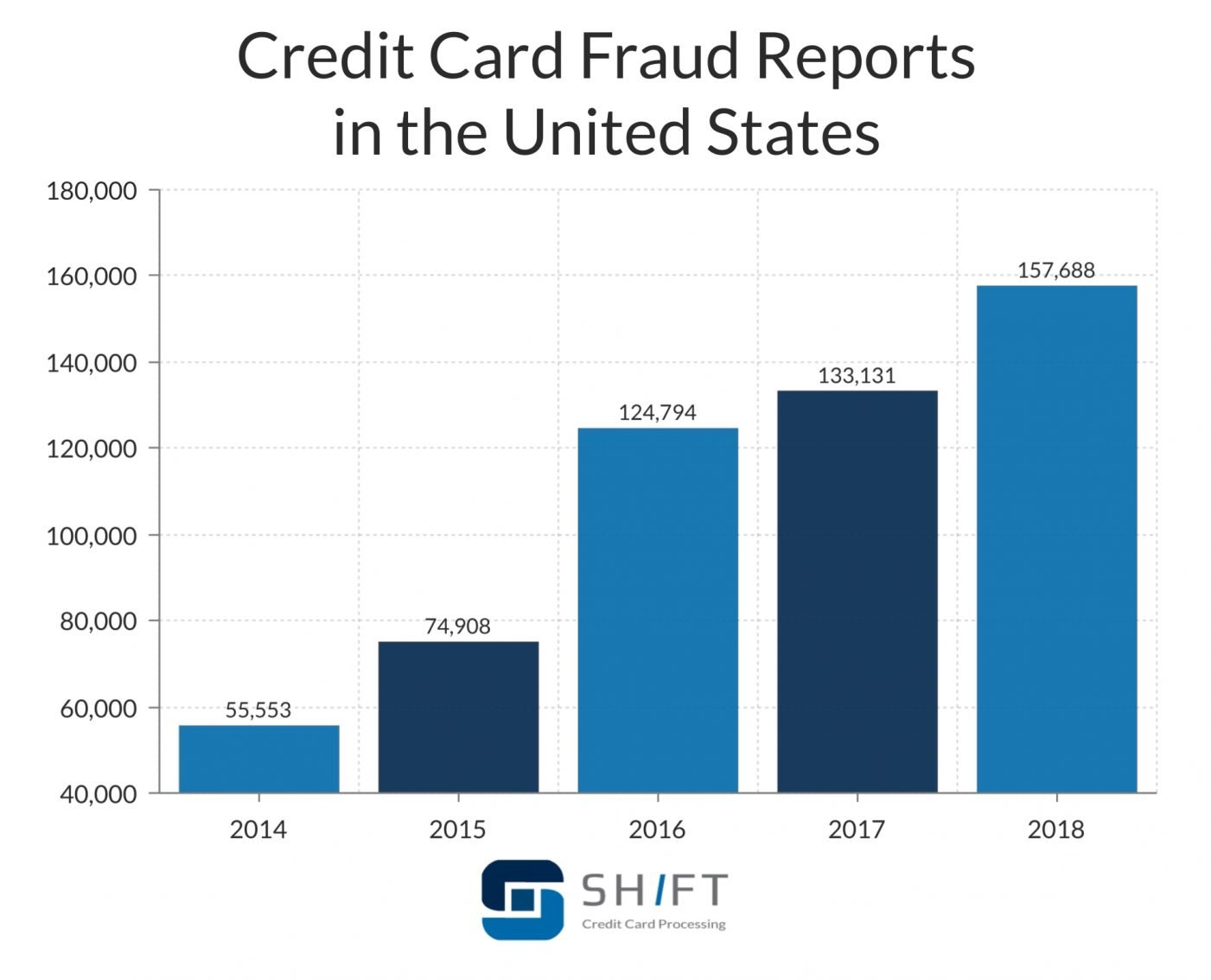
Credit cards and electronic payments make overall functioning in a global marketplace much easier. Credit cards offer ease of transaction at merchant outlets, allowing an individual or businesses to purchase things and pay later. Credit cards offer you a line of credit that can be used to make purchases, balance transfers and/or cash advances and requiring that you pay back the loan amount in the future. As a part of the economic growth, consumers can borrow and spend more whereas enterprises can borrow and invest more. A rise of consumption and investments creates jobs and leads to a growth of both income and profit. A rise of consumption and investments creates jobs and leads to a growth of both income and profit.

If one wants to fully appreciate the modern convenience of credit cards, simply insert your chip card, pause while it processes, and consider what all it has replaced. Cash works best in face-to-face transactions and checks take time to cash. Credit cards are not only important for individuals and businesses, but they are also a very important aspect of continued economic growth.



**Problem Statement**

Each year financial institutions lost a chunk of money as a result of credit card fraud. In year 2018, a total of $24.26 Billion was lost due to payment card fraud across the globe and United States being the most fraud prone country. Credit card fraud was ranked number one type of identity theft fraud. Credit card fraud increased by 18.4 percent in 2018 and is still climbing. Credit card fraud includes fraudulent transactions on a credit card or debit card. There can be two kinds of card fraud, card-present fraud and card-not-present fraud. Card not present fraud is almost 81 percent more likely than point-of-sale fraud.



It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase. This would not only result in financial loss but also loss of customer confidence in payment industry.

**Scope**

Credit card transactions data is highly sensitive information, therefore the features in the dataset are scaled, and the names of the features are not shown due to privacy reasons. There is still a lot of relevant information that we still can use to analyze some important aspects of the dataset. The scope of this project would be to create a prediction model to detect credit card fraud based on the features in this dataset.

“The dataset has been collected and analyzed during a research collaboration of Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection.”

**Document Overview**

This proposal is broken down into five sections: Introduction, Preliminary Requirement, Technical Approach, Expected Results, and Management Approach. Many of these major sections will contain sub-sections that will show the tables, figures, and analysis that went into the predictive model.

**Preliminary Requirement**

**Technical Approach**

The project will be carried out by utilizing the CRISP-DM model. It stands for Cross Industry Standard Process for Data Mining. The process contains 6 steps that will be followed throughout the project.

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

During every phase of this project lifecycle, we might discover new aspects/finding which we will incorporate them in ways to improve the efficiency of our model.

**Data Sources and Variables**

The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. It contains only numerical input variables which are the result of a PCA transformation.

Attribute Information:

1. Time - Number of seconds elapsed between this transaction and the first transaction in the dataset.
2. V1- V28 – These are the result of a PCA Dimensionality reduction to protect user identities and sensitive features.
3. Amount – Transaction amount
4. Class – This is a response variable and has the values of 1 for fraudulent transactions, and 0 for non-fraudulent transactions.

There are 29 decimal fields and 2 integer fields in the dataset.

Dataset can be found on Kaggle: <https://www.kaggle.com/mlg-ulb/creditcardfraud>

**Analysis**

Our dataset contains 31 different variables that could contribute to our predictive model. We will perform feature reduction and dimensionality reduction to select the most relevant variables for our model. According to our text, “Applied Predictive Analytics”, many predictive algorithms assume the model variables follow a normal distribution. There are inherent advantages to using normally distributed variables, so our approach will focus on columns that closely follow this distribution.

We will determine which variable are normally distributed by conducting the following analyses:

* Summary statistics on all variables—concentrating on mean and standard deviation values
* Skewness of variables—looking for variables with values less than two and as close to zero as possible

**Requirement Development**

We plan to rely almost exclusively on Jupyter Lab, building a notebook leveraging the Python language. Based on our requirements, we plan to download and install necessary packages to achieve desired results. If any R coding is required, we always can add that into the same notebook via available Python libraries.

**Model Deployment**

With the help of graph analysis, we will understand the dataset and determine some of the important variables which can be used for our models. We will use feature selection techniques to finalize our feature list for models. Using the results from this step, we will build a couple of classification models and evaluate its performance. We will implement ways on how we can improve our model efficiency.

**Testing and Evaluation**

We will be using the train-test split of 70%-30% and will test the model with test split. For evaluation confusion matrix, AUC, and F1 score will be used. For some models cross-validation will also be used to decide best model.

**Expected Results**

The objective is to build an efficient model to determine if a credit card transaction is fraudulent or not.

**Execution and Management of Project**

**Project Plan**

We reviewed the project milestones of this course and believe that those milestones are the best way to move forward for the timely completion of our project.

* Week 1: Team Information/Communication Plan
* Week 2: Data Selection and Project Proposal
* Week 3: Data set cleanup
* Week 4: Graph Analysis & EDA
* Week 5: Peer review, discussion & finalize week 3 & 4 work
* Week 6: Perform feature selection review different classification models
* Week 7: Train and test the models
* Week 8: Find ways to improve model efficiency
* Week 9: Merge and finalize for Milestone 4 submission
* Week 10: Review feedback & adjust accordingly
* Week 11: Work on documentation & presentation
* Week 12: Final project paper and presentation

**Project Risk**

One of the biggest challenges would be to find the most relevant features for our classification model. We don’t have the visibility of the type of data in columns V1-28 which may pose few challenges while we are trying to improve the efficiency of the model.

**Resources**

ULB, M. (2018, March 23). Credit card fraud detection. Retrieved March 27, 2021, from <https://www.kaggle.com/mlg-ulb/creditcardfraud>

Credit card Fraud Statistics. (2021, January 04). Retrieved March 27, 2021, from <https://shiftprocessing.com/credit-card-fraud-statistics/>