1. **Credit Card Fraud Detection**

This project is intended to detect fraudulent transactions and build a predictive model to detect fraudulent transactions as it occurs or identify suspect after fraudulent transaction has occurred.

**Problem-**

Build machine learning model using dataset that contains fraudulent transactions made by credit cards in September 2013 by European cardholders. This model shall be used to predict fraudulent transactions as they occur.

**Client-**

Credit Card companies

Why do they care about the problem?

According to some estimates one out of every 1200 transactions are fraudulent ([Wikipedia Link](https://en.wikipedia.org/wiki/Credit_card_fraud#cite_note-Fraud_is_a_rare_event-3)), this amounts to millions of dollars. Credit card companies shouldn’t be charging the customer for transaction they didn’t perform.

**Data**

<https://www.kaggle.com/mlg-ulb/creditcardfraud>

The datasets contain 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. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependent cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

**Modeling Approach**

Logistic regression model will be used to predict fraudulent transaction.

Coefficients of the Logistic regression will be estimated from the training data using maximum-likelihood estimation. The coefficients result in a value close to 1(in case of fraud) and 0 (in case of non-fraud).

Model will first be trained on train data and will be validated against the test data to accuracy of the model.

**Deliverables**

1. Code (Jupyter notebook)
2. Report on finding of the project
3. Slide deck
4. **Medicare Data**

This project intends to uncover trends in the medicare data using machine learning algorithms.

**Problem-**

Charge variation with respect to item category, procedure and service, physician and location?

What is the total number of medications prescribed in each state?

What is the most prescribed medication in each state?

What is the average cost for inpatient and outpatient treatment in each city and state?

Which are the most common inpatient diagnostic conditions in the United States?

Which cities have the most number of cases for each diagnostic condition?

What are the average payments for these conditions in these cities and how do they compare to the national average?

**Client-**

Centers for Medicaid and Medicare Services and beneficiaries of Medicare.

Why do they care about the problem?

Analysis of data and trend will help to better focus the program for its beneficiaries. Any possible misuse of the program.

**Data**

[**https://www.kaggle.com/cms/cms-medicare**](https://www.kaggle.com/cms/cms-medicare)

[**https://cloud.google.com/bigquery/public-data/medicare**](https://cloud.google.com/bigquery/public-data/medicare)

Medicare Data (BigQuery Dataset) – Not Sure how to access data for this

This public dataset was created by the Centers for Medicare & Medicaid Services. The data summarizes the utilization and payments for procedures, services, and prescription drugs provided to Medicare beneficiaries by specific inpatient and outpatient hospitals, physicians, and other suppliers. The dataset includes the following data.

Common inpatient and outpatient services All physician and other supplier procedures and services All Part D prescriptions. Providers determine what they will charge for items, services, and procedures provided to patients and these charges are the amount that providers bill for an item, service, or procedure.

**Modeling Approach**

Linear regression, using clustering and classification trees approach will be used to analyze trends on medicare data.

**Deliverables**

1. Code (Jupyter notebook)
2. Report on trends, analysis and finding of the project
3. Slide deck
4. **Amazon Reviews for Sentiment Analysis**

This project intends to provide sentiment analysis of the customer.

**Problem:**

Sentiment Analysis

1. What features of the product are mostly liked by the customers?
2. What the improvement areas in the product?
3. What do customer feel about the customer service?
4. What are the mostly commonly used words by the customer to express their feedback?

**Client-**

Analysis of the reviews help the companies to understand the how customer feedback of their product.

**Data:**

[**https://www.kaggle.com/bittlingmayer/amazonreviews**](https://www.kaggle.com/bittlingmayer/amazonreviews)

A few million Amazon reviews in fastText format

\_\_label\_\_<X> \_\_label\_\_<Y> ... <Text>

where X and Y are the class names. No quotes, all on one line.

In this case, the classes are \_\_label\_\_1 and \_\_label\_\_2, and there is only one class per row.

\_\_label\_\_1 corresponds to 1- and 2-star reviews, and \_\_label\_\_2 corresponds to 4- and 5-star reviews.

(3-star reviews i.e. reviews with neutral sentiment were not included in the original),

The review titles, followed by ':' and a space, are prepended to the text.

Most of the reviews are in English, but there are a few in other languages, like Spanish.

**Modeling Approach**

Naïve Bayes Classifier approach will be used for sentimental analysis.

Words that express emotions shall be identified and shall be categorized by their emotions. This classification shall be used on the dataset identify sentiment.

**Deliverables**

1. Code (Jupyter notebook)
2. Report on trends, analysis and finding of the project
3. Slide deck