

CMP4011 Big Data and Cloud Computing

Project Proposal

Team 9

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Sec** | **B.N** | **Code** |
| Ahmed Hosny | 1 | 2 |  |
| Ahmed Sabry | 1 | 4 |  |
| Basma Elhoseny | 1 | 16 | 9202381 |
| Zeinab Moawad Fayez | 1 | 28 |  |

Idea (1) Credit Score Classification Problem

## Problem Statement:

To assess the risk associated with approving a loan request from a customer, bankers rely on evaluating the customer's transactional history. We propose the development of an intelligent system that would analyze customer credit information, encompassing payment history, credit utilization, and length of credit. It would then assign a credit score to the customer, which the bank would use to determine loan approval based on a predefined threshold.

## Questions we will Answer:

* What is the Segment of this Customer? [Poor-Standard-Good]
* How is Factors such as age, occupation is correlated with monthly salary and average balance

## Data Set:

• **Main:** <https://www.kaggle.com/datasets/parisrohan/credit-score-classification> [31.14 MB 150K Example] [25 useful feature excluding name, id, ssn, and month]

• <https://www.kaggle.com/datasets/clkmuhammed/creditscoreclassification?select=train.csv> [Cleaner Version but we think we need to achieve this clean version from the one above it is a step to be done by us 😊)

• <https://www.kaggle.com/code/jayrdixit/credit-scoring> [Unsupervised]

• <https://www.kaggle.com/datasets/conorsully1/credit-score> [Small Data set but with huge # of features with regression as score but we need to check with TA if we can make use of it with our main dataset]

## Approach:

##### EDA (exploratory Data Analysis) Phase:

1. Carry out Statistical Analysis on the data set computing mean std ……
2. Anomalies and outliers Detections
3. Plotting Distributions [Data Visualization]
4. Data Cleaning and Handling missing values.
5. Checking correlations between features [Correlation Analysis]
6. We may need feature space reduction as PCA [to be checked later when we start the analysis phase]

##### Descriptive Analysis Methods:

1. KMeans Clustering to segment customers into clusters based on their credit scores.

##### Predictive Analysis Methods

1. Random Forest
2. KNN
3. Naïve Baye’s Classifier

We Think of Implementing **Random Forest** or K means Clustering by Map Reduce Approach.

Idea (2) Patient Stay Length Problem

## Problem Statement:

For a hospital trying to improve its health care service the patient stay period is a very critical parameter, It helps hospitals to identify patients of high LOS risk (patients who will stay longer) at the time of admission. Once identified, we can adjust the treatment plan to be optimized in order to minimize LOS and lower the chance of staff/visitor infection. Also, prior knowledge of LOS can aid in logistics such as room and bed allocation planning.

## Questions we will Answer:

* How Long will this patient probably stay? Categorial Answer [11 different classes ranging from 0-10 days to more than 100 days]
* How does the illness severity along with the stay period affect admission value

## Data Set:

• **Main:** <https://datahack.analyticsvidhya.com/contest/janatahack-healthcare-analytics-ii/True/#ProblemStatement\> [31.14 MB 318K Example] [16 useful feature excluding name and id]

## Approach:

##### EDA (exploratory Data Analysis) Phase: [Same as Idea (1)]

1. Carry out Statistical Analysis on the data set computing mean std ……
2. Anomalies and outliers Detections
3. Plotting Distributions [Data Visualization]
4. Data Cleaning and Handling missing values.
5. Checking correlations between features [Correlation Analysis]
6. We may need feature space reduction as PCA [to be checked later when we start the analysis phase]

##### Descriptive Analysis Methods:

1. KMeans Clustering to segment customers into clusters based on their credit scores.

##### Predictive Analysis Methods

1. Random Forest
2. KNN
3. Naïve Baye’s Classifier

We Think of Implementing **Naïve Bayes** or K means Clustering by Map Reduce Approach.

Idea (3) Vehicle Sales

## Problem Statement:

Are you thinking of buying your car and even buying a second-hand car? But what is the price you expect to pay? Several factors affecting car prices such as model, year, odometer reading. Using this data, we can build insights about these factors and build a predictive model to predict the average price for the car with the given specs.

## Questions we will Answer:

* What is the Estimated value given its specifications data.
* How Is the sessional date affecting price of the car.

## Data Set:

• **Main:** <https://www.kaggle.com/datasets/syedanwarafridi/vehicle-sales-data> [88MB 550K examples] [16 useful features]

## Approach:

##### EDA (exploratory Data Analysis) Phase: [Same as Idea (1)]

1. Carry out Statistical Analysis on the data set computing mean std ……
2. Anomalies and outliers Detections
3. Plotting Distributions [Data Visualization]
4. Data Cleaning and Handling missing values.
5. Checking correlations between features [Correlation Analysis]
6. We may need feature space reduction as PCA [to be checked later when we start the analysis phase]

##### Descriptive Analysis Methods:

1. KMeans

##### Predictive Analysis Methods

1. Linear Regression
2. Logistic Regression

We Think of Implementing **Random Forest** or K means Clustering by Map Reduce Approach.

DRAT

## Healthcare Analytics II

<https://datahack.analyticsvidhya.com/contest/janatahack-healthcare-analytics-ii/True/#ProblemStatement\>

1. ~~EDA (exploratory data analysis) Phase:~~
2. ~~Carry out statistics on the data set like computing mean std …… Plotting Distributions [Data Visualization]~~
3. ~~Data Cleaning and Handling missing values like using men and std.~~
4. ~~Checking correlations between features [~~**~~Correlation Analysis]~~**
5. **~~Reduction of feature space using techniques like PCA~~**

The algorithms that we think if implementing using Map Reduce Algorithm

1. Clustering Techniques:
   1. Case Classification 🡪 KNN
   2. Case Regression 🡪 KMeans
2. Random Forest for both regression and categorical 🡪 multiple decision trees and combines their predictions through
   1. Case Classification 🡪 voting.
   2. Case Regression 🡪 Aggregate these predictions (e.g., by averaging) to obtain the final regression prediction.
3. **Naive Bayes**: For Categorial Classification Problems
4. **Logistic Regression: For Classification & Regression Problem**

The descriptive analysis methods to be used:

1. Clustering Using Kmeans

The predictive analysis methods: