CMP4011 Big Data and Cloud Computing

Project Proposal

Team 9

Name	Sec	B.N	Code
Ahmed Hosny	1	2	9202077
Ahmed Sabry	1	4	9202119
Basma Elhoseny	1	16	9202381
Zeinab Moawad Fayez	1	28	9202611

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 are factors such as age and occupation 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]

Approach:

EDA (exploratory Data Analysis) Phase:

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

Descriptive Analysis Methods:

i. KMeans Clustering to segment customers into clusters based on their credit scores. [Further Clustering]
   a. This idea is insighted from https://www.kaggle.com/code/jayrdixit/credit-scoring [Unsupervised]
ii. Association Rules Between Features
   a. Such as {Occupation Doctor} => {Credit Score High}.
   b. Such as {Occupation Developer} => {Credit Score Low}.

Predictive Analysis Methods

i. Random Forest [**Map Reduced**]
ii. KNN [**Map Reduced**]
iii. Naïve Baye's Classifier [**Map Reduced**]
iv. Logistic Regression
v. SVM

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 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)]

i. Carry out Statistical Analysis on the data set computing mean std ......
ii. Anomalies and outliers Detections
iii. Plotting Distributions [Data Visualization]
iv. Data Cleaning and Handling missing values.
v. Checking correlations between features [Correlation Analysis]
vi. We may need feature space reduction as PCA [to be checked later when we start the analysis phase]

Descriptive Analysis Methods:

i. KMeans for segmenting patients into clusters each cluster for different duration. [Further Segmentation]
ii. Association Rules Such as:
    a. {Severity of Illness} => {Stay}
    b. {Ward Type} => {Severity of Illness}

Predictive Analysis Methods

i. Random Forest [**Map Reduced**]
ii. KNN [**Map Reduced**]
iii. Naïve Baye's Classifier [**Map Reduced**]
iv. Logistic Regression
v. SVM

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)]

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

Descriptive Analysis Methods:

i. KMeans
   a. Segmenting Vehicles into Clusters Based on features other than selling price.
   b. Clustering for the selling price so to have categorial feature to be used later.
ii. Association Rules such as:
   a. {Transmission Type = Automatic, Body Type = Sedan} => {Selling Price = High}
   b. {Transmission Type = Manual, Body Type = Truck} => {Condition = Used}

Predictive Analysis Methods

i. Linear Regression
ii. Support Vector Regression
iii. Ridge Regression