Assignment 1 – Missing Values

Praveen Mahaulpatha (C0860583)

Artificial Intelligence and Machine Learning

AML 1114 2 – Data Science and Machine Learning

Debashish Roy

July 11, 2022

Part 1 : Understanding of Data Science and Missing Values

A. Define data science. Describe how data science can be usable in the following sectors (use any two areas of your choice)

Data science is the process of encompassing a set of principles, problem definitions, algorithms, and processes for extracting non obvious and useful patterns from large datasets.(John D. Kelleher & Brenden Tierney, 2018) A data scientist’s duties typically include strategy development for data analysis, preparation of data for analysis, exploring, analyzing, and visualizing data, model building with data using programming languages, such as Python and R, and deploying models into applications. (Oracle, 2022)

a. Banking and Financial Sector

The use of Artificial Intelligence and Data Science(AIDS) has been in rise for Economics and Finance(EcoFin) sector. The main financial assets, products, instruments, and their related operations and services that can benefit from AIDS include the following.(Longbing Cao, 2022)

- Stock and services
- Derivative and services
- Commodities and services
- Index and services
- Currency, cryptocurrency, and services

- Banking and services
- Insurance and services
- Wealth and services
- Surveillance and compliance

Some of the common implications of Artificial Intelligence and Data Science are (Arthur Bachinskiy, 2019)

1. Aiding Credit decisions:

AIDS allows accurate assessment of a borrower with powerful algorithms and large volume of data. Banks use machine learning to evaluate loan eligibility by analysing more parameters faster. This has helped apps which uses auto lending features to decide low risk borrowers.

2. Risk management:

Large processing powers of modern computers coupled with efficient machine learning algorithms helps to analyze immense volumes of structured and unstructured data faster to identify potential risks.

3. Fraud prevention:

AIDS is most commonly used in credit card fraud detection. This sector has been growing faster than ever due the increased number of online transactions taking place. The systems analyze clients’ behavior, location, and buying habits and trigger a security mechanism when something seems out of order and contradicts the established spending pattern.

4. Trading:

Intelligent trading systems (Trading Bots) allows users to make automatic trading with minimum risk and high precision. It allows users to analyze vast amount of trading data to predict market behaviour and identify potential opportunities for undervalued stocks. The system provides solution across different requirements including well diversified portfolios for long-term and short-term investments.

5. Personalized banking:

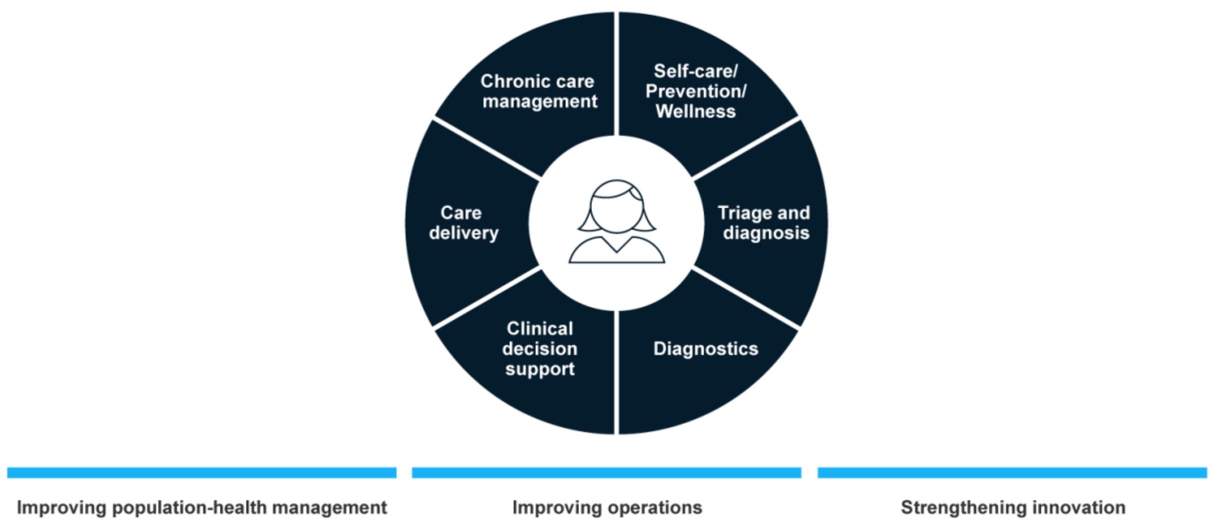
AI powered chatbots and voice-controlled assistance makes it easy to cater for individual customer needs more efficient and faster. Many of the apps supported by data science track income, essential recurring expenses, and spending habits and come up with an optimized plan and financial tips.

6. Process automations:

Machine learning enables computers to learn recursive patters and to learn to do it on their own. This saves human capital but also reduces the chances of human error with robust models. Systems can review, analyze, extract information and generate reports faster than ever with minimum human interventions.

b. Healthcare

Areas of impact for AI in healthcare.



McKinsey & Company

Above figure shows the sectors of healthcare where AI and ML are commonly used. As the data scientist now have access to immense number of data records from medical institutes, it has made possible to derive patterns and predict or even prescribe what decisions to make. Powerful deep learning techniques such as convolutional neural networks for image processing has enabled systems to detect potential health risks by given an image of a X ray or a scan, with better accuracy than a human specialist. Other use cases are smart organizing scheduling or bed management systems, predicting hospital admission rates and accelerating R&D for new treatments. (Spatharou et al., 2022)

B. What are the missing values and errors in data?

Missing values are defined as values that are not stored or having empty values in a given dataset. Usually it is represented as blank in datasets and in pandas as NaN. There can be several reasons for missing values including sampling error, missing to add data, corrupted data, etc. There are three types of missing values; Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR).

Part 2 : Experiment on Missing Values:

2.1 Project Overview

Project Purpose	Predicting whether a Meteorite is hazardous or not using meteorite characteristics.
IDE	Jupyter Notebook
Python Version	Python 3.9
Machine Learning Technique	Supervised Learning
Model	Classification
Dataset Source	Kaggle
Predictor Variables	34
Response Variable	1
Libraries	Pandas / Numpy / matplotlib / sklearn

2.2 Dataset

The dataset is a CSV collection of data with total of 40 rows and 4,687 rows. It has below features and corresponding data types.

```
Neo Reference ID                    int64
Name                                int64
Absolute Magnitude                float64
Est Dia in KM(min)                float64
Est Dia in KM(max)                float64
Est Dia in M(min)                 float64
Est Dia in M(max)                 float64
Est Dia in Miles(min)             float64
Est Dia in Miles(max)             float64
Est Dia in Feet(min)              float64
Est Dia in Feet(max)              float64
Close Approach Date                object
Epoch Date Close Approach         float64
Relative Velocity km per sec      float64
Relative Velocity km per hr       float64
Miles per hour                    float64
Miss Dist.(Astronomical)          float64
Miss Dist.(lunar)                 float64
Miss Dist.(kilometers)            float64
Miss Dist.(miles)                 float64
Orbiting Body                      object
Orbit ID                            int64
Orbit Determination Date           object
Orbit Uncertainity                  int64
Minimum Orbit Intersection        float64
Jupiter Tisserand Invariant       float64
Epoch Osculation                  float64
Eccentricity                      float64
Semi Major Axis                   float64
Inclination                       float64
Asc Node Longitude                float64
Orbital Period                    float64
Perihelion Distance               float64
Perihelion Arg                    float64
Aphelion Dist                     float64
Perihelion Time                   float64
Mean Anomaly                      float64
Mean Motion                       float64
Equinox                            object
Hazardous                            bool
```

The feature “Hazardous” which is represented in bool is the response variable which we try to predict. For the problem we have removed features containing ‘object’ datatypes as it cannot be computed for selected statistical models.

2.3 Steps taken to handle Missing Values

a. Check the number of missing values for each feature

```python
#Check for missing values
df.isnull().sum()
```

```
Neo Reference ID                  0
Name                              0
Absolute Magnitude                0
Est Dia in KM(min)                0
Est Dia in KM(max)                0
Est Dia in M(min)                35
Est Dia in M(max)                35
Est Dia in Miles(min)            35
Est Dia in Miles(max)            35
Est Dia in Feet(min)             35
Est Dia in Feet(max)              0
Close Approach Date               0
Epoch Date Close Approach        10
Relative Velocity km per sec      0
Relative Velocity km per hr       0
Miles per hour                    0
Miss Dist.(Astronomical)          0
Miss Dist.(lunar)                 0
Miss Dist.(kilometers)           23
Miss Dist.(miles)                 0
Orbiting Body                     0
Orbit ID                          0
Orbit Determination Date          0
Orbit Uncertainity                0
Minimum Orbit Intersection        0
Jupiter Tisserand Invariant       0
Epoch Osculation                  0
Eccentricity                      0
Semi Major Axis                   0
Inclination                       0
Asc Node Longitude                0
Orbital Period                    0
Perihelion Distance               0
Perihelion Arg                    0
Aphelion Dist                     0
Perihelion Time                   0
Mean Anomaly                      0
Mean Motion                       0
Equinox                           0
Hazardous                         0
```

b. Remove missing values

Dropping rows with missing values of the dataset.

```python

df.dropna(inplace=True)

```

c. Imputing missing values with mean

Filling the mean value of correspondent column for missing values.

```
df.fillna(df.mean(), inplace=True)
```

Furthermore we have experimented with below steps:

d. Fitting missing value non supportive model

Linear Discriminant Analysis is a classification model used in supervised learning which does not support missing values. The experiment was carried out to test this scenario and the result output was a value error with below message:

```
ValueError: Input contains NaN, infinity or a value too large for dtype('float64')
```

e. Fitting LDA after removal of missing values

Since LDA does not support missing values, first we have removed the missing values and tried fitting the model and calculated the accuracy using Kfold cross validation method.

Ouput Accuracy:

```
Accuracy: 0.916
```

f. Fitting LDA after imputing missing values with mean

Secondly, the missing values were imputed with mean values and fitted to LDA model to check the accuracy using Kfold cross validation method.

Output Accuracy:

```
Accuracy: 0.920
```

It is evident that by imputing we have gained a better accuracy than removal of missing values.

g. Fitting other Classification models to compare accuracy

In this section we have fitted the mean imputed dataset to 5 different models and compared their accuracy and standard deviation using the Kfold cross validation. The below output was given for each model.

```
LR: 0.838936 (0.029507)
LDA: 0.918088 (0.025396)
KNN: 0.838936 (0.029507)
CART: 0.995521 (0.006568)
NB: 0.838936 (0.029507)
SVM: 0.838936 (0.029507)
```

It was evident that the Decision Tree Classifier(CART) model was yielding the best results for the dataset we have at hand with a 99.5% of accuracy and a very low standard deviation.

h. Feature Importance

Checking the feature importance for a selected model helps to understand the most influential feature for the prediction. We can use the 'feature_importances_' method on the model to do so. Below is the output we received for the Decision Tree Classifier in descending order.

features	feature_importance
Est Dia in KM(max)	0.785114
Minimum Orbit Intersection	0.185483
Est Dia in Feet(min)	0.005776
Neo Reference ID	0.005187
Absolute Magnitude	0.004306
Epoch Date Close Approach	0.003902
Orbit ID	0.002766
Miss Dist.(lunar)	0.002456
Perihelion Distance	0.002367
Asc Node Longitude	0.001381
Eccentricity	0.000714
Perihelion Time	0.000474
Est Dia in KM(min)	0.000073
Perihelion Arg	0.000000
Aphelion Dist	0.000000
Orbital Period	0.000000
Inclination	0.000000
Semi Major Axis	0.000000
Mean Anomaly	0.000000
Epoch Osculation	0.000000
Jupiter Tisserand Invariant	0.000000
Miss Dist.(kilometers)	0.000000
Orbit Uncertainity	0.000000
Miss Dist.(miles)	0.000000
Name	0.000000
Miss Dist.(Astronomical)	0.000000
Miles per hour	0.000000
Relative Velocity km per hr	0.000000
Relative Velocity km per sec	0.000000
Est Dia in Feet(max)	0.000000
Est Dia in Miles(max)	0.000000
Est Dia in Miles(min)	0.000000
Est Dia in M(max)	0.000000
Est Dia in M(min)	0.000000
Mean Motion	0.000000

It is evident that the most important feature is the 'Est Dia in KM(max)' which is the estimated diameter of the meteorite with a 78% importance.

2.4 Conclusion

In this report we have emphasized the importance of dealing with missing values in machine learning before fitting statistical models. We have can mainly deal with missing values either by dropping them altogether or imputing them with a different value (mean is preferred). The model predicts with different accuracy levels for each of these scenarios. Once that was clear, we have fitted different other classification models and understood the best result was given by Decision Tree Classifier. Later the most significant feature was identified as the diameter of meteorite. The accuracy can further be increased of the model by checking for outliers, normalizing data and other data cleansing methods.

References

Data Science. 2018. John D. Kelleher & Brenden Tierney. MIT Press

What is Data Science?. 2022. Oracle

https://www.oracle.com/ca-en/data-science/what-is-data-science/

AI in Finance: Challenges, Techniques, and Opportunities. (03 February 2022). Longbing Cao

The Growing Impact of AI in Financial Services: Six Examples. (21 Feb 2019). Arthur Bachinskiy

Transforming healthcare with AI: The impact on the workforce and organizations. (2022). Angela Spatharou, Solveigh Hieronimus, and Jonathan Jenkins.