

**NAME**

**COLLEGE NUMBER**

**TITLE**

**Design and Application of a Machine Learning System for a Practical Problem:**

**A case study of the hotel industry openings in new locations**

## INTRODUCTION

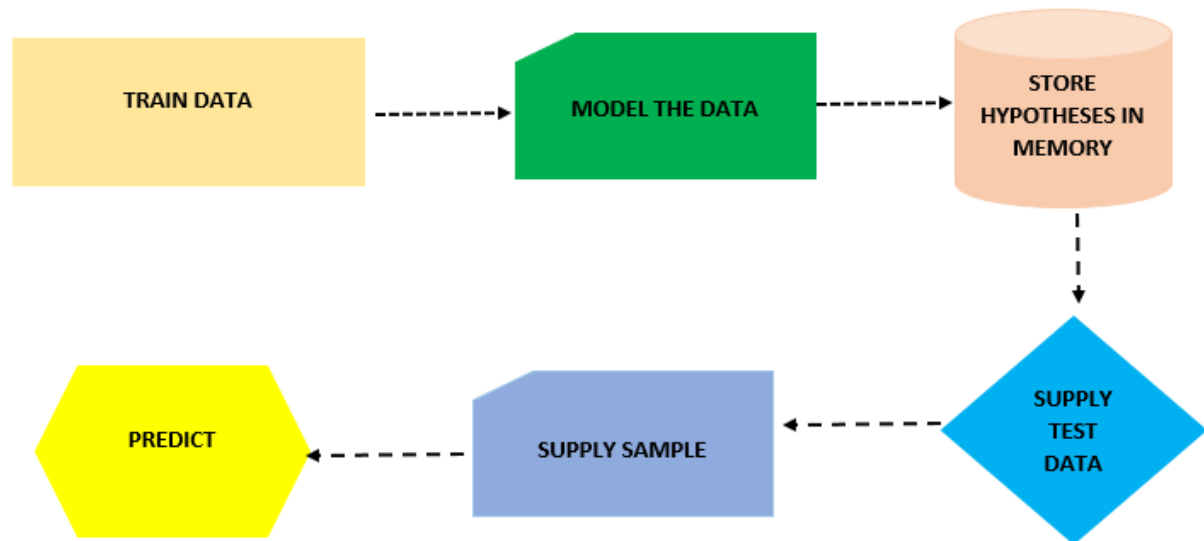
In a fast growing world and economy, organisations are finding the need to sustain data a more crucial role within their departments and other cross functional applications. Further decision support tools are not just enough to satisfy the urge that management has towards fulfilling some of the key needs that they are faced with when it comes to delay finding just the perfect tool to help them actualise decision rules. Customer relationship management systems, enterprise resource planning tools, ERP portals, knowledge management systems and bases, social media platforms and pages, are some of the avenues through which organisations acquired data.

Since big data is that which is fast moving, in large volumes, is not organised and is in many forms, there is the hidden potential behind this data that it could be carrying some of the greatest mines that need to be harvested and projected for organisational performance and maintenance. Tools such as Hadoop and Apache Spark have been used before to ensure that data is mined, cleaned and trimmed for further analysis.

Further to this process, data centre infrastructure is another key proponent that needs to be taken into key consideration by the various system engineers to ensure that there are enough physical infrastructural facilities to ensure that data is properly centred, through the various data entry points such as social media platforms, company forms and contact points and all lead generators. Once this has been done, historical partitioning of the database systems need to be done and established in a manner that can be accessed by the various data stakeholders.

In addition, Business intelligent systems are a more refined way of getting reliable decisive information. BI systems filter the data, summarise it and then present them in clear graphs and presentational formats that all managers and top decision makers can consume and act upon. Apart from just summarising and reporting the data, organisations need predictive analytics as part of their key decision making metrics. Meaning that, there is always the need to project and try to define what the future might look like, this depends on the application of machine learning algorithms to the dataset and then applying the relevant models to the data so that the predictive functions can be applied to the selected dataset. Usually the A/B test rule will be applied to the data where 80% of the data shall be used for training and then 20% for testing

purposes. During, the training, the machine shall learn from this set and then testing the possible outputs on the result. An illustration is given below:



## **Problem statement**

In this exercise, the researcher established the potential of investing in a new store by trying to establish whether investing a new shop in a new location would be viable. In order to achieve this, a decision tree algorithm was supplied to the dataset and the sample data of variable population and purchase revenues used for testing, the resultant data was then predicted to establish the Boolean of TRUE for profitable and FALSE for not profitable

## **Project objectives and aims**

- To learn to identify machine learning techniques appropriate for establishing the viability of a new store in a new location
- To undertake a comparative evaluation of several machine learning procedures when applied to the specific problem
- To predict the possibility of a new hotel branch being profitable in a new town

## Research methodology and design

The dataset supplied and used for this particular analyses was obtained from the store data centres that contained information about the various store locations and sales plus population information about these store locations. A quick summary this dataset looks like below.

| F1       | F2      | F3       | F4       | F5   | F6       | F7      | F8       | F9      | F10      | F11      | F12     | F13     | F14 | F15     | F16      | F17    | F18     | F19     | F20 | F21   |
|----------|---------|----------|----------|------|----------|---------|----------|---------|----------|----------|---------|---------|-----|---------|----------|--------|---------|---------|-----|-------|
| -0.12321 | 2577.32 | 2183.52  | 7747.32  | 2.16 | 9693.21  | -2.2498 | -1381.98 | 3033.55 | 53.796   | -90073.7 | 2.2986  | 6.786   | 1   | -3.929  | -0.8095  | 3.58   | -5.9384 | -9.342  | 1   |       |
| -2.0646  | 1200.12 | 1654.74  | 6027.72  | 2.16 | 9391.26  | -2.742  | -1336.2  | 2081.35 | 33.514   | -90039.9 | 0.9676  | 5.524   | 0   | -1.6609 | -2.302   | 3.888  | -6.564  | -6.3806 | 1   |       |
| -0.08375 | 2769.35 | 2661.42  | 7798.66  | 2.16 | 9296.28  | -3.213  | -1531.16 | 3397.15 | 39.43    | -90138.3 | 1.7296  | 10.087  | 0   | -2.51   | -2.819   | 2.1606 | -5.234  | -9.678  | 0   |       |
| -2.8083  | 3161.52 | 2199.12  | 8369.32  | 5.16 | 7990.71  | -2.2703 | -1397.03 | 3289.99 | 42.012   | -90993.3 | 3.532   | 15.54   | 1   | -2.2555 | -0.977   | 8.216  | -5.8166 | -6.1512 | 0   | 42.06 |
| -0.4458  | 2707.42 | 2083.62  | 9054.72  | 5.16 | 8691.81  | -3.731  | -804.04  | 1903.35 | 43.172   | -90092.3 | 0.5662  | 4.5188  | 0   | -1.7843 | -1.369   | 0.7562 | -6.406  | -9.532  | 0   |       |
| -4.638   | 2716.72 | 2248.02  | 12806.92 | 5.16 | 12720.21 | -1.9979 | -1736.04 | 4614.35 | 41.964   | -90068   | 0.54782 | 11.924  | 0   | -1.4422 | -0.4898  | 1.2142 | -6.2044 | -7.3166 | 1   |       |
| -0.29571 | 2792.12 | 1338.12  | 7499.32  | 2.16 | 9608.91  | -1.7279 | -1464.12 | 3537.65 | 34.556   | -89874.2 | 1.1074  | 32.38   | 1   | -2.565  | -1.1173  | 3.638  | -5.7458 | -8.006  | 0   |       |
| -0.4449  | 1891.42 | 3179.82  | 8664.32  | 5.16 | 9398.88  | -2.2642 | -1353.14 | 3015.15 | 47.606   | -90117.3 | 1.2584  | 13.707  | 0   | -1.9235 | -0.7047  | 1.5188 | -4.8998 | -6.1578 | 1   | 40.5  |
| -1.2639  | 2477.32 | 2659.92  | 7834.056 | 5.16 | 10069.71 | -3.137  | -1413.68 | 3038.75 | 33.852   | -90313.1 | 1.6374  | 14.89   | 0   | -3.928  | -0.7701  | 1.846  | -5.106  | -7.786  | 0   |       |
| -0.16245 | 2914.19 | 1502.7   | 9395.32  | 2.16 | 8930.52  | -2.4263 | -664.44  | 3268.76 | 30.058   | -90078.6 | 2.1702  | 5.604   | 0   | -1.561  | -2.918   | 0.5416 | -5.598  | -7.792  | 1   | 40.92 |
| -0.06085 | 2705.52 | 1513.47  | 9075.32  | 2.16 | 10616.01 | -3.073  | -1738.24 | 3258.56 | 39.696   | -90132.9 | 1.1492  | 9.079   | 0   | -1.896  | -0.6148  | 0.4914 | -5.1992 | -6.777  | 1   |       |
| -1.7295  | 2648.32 | 2289.72  | 7706.42  | 5.16 | 11583.81 | -2.575  | -1706.44 | 5928.35 | 29.21    | -90133.7 | 0.50936 | 5.621   | 1   | -4.387  | -0.6011  | 1.6988 | -5.6096 | -6.9204 | 0   |       |
| -5.22    | 691.12  | 1786.98  | 7647.82  | 5.16 | 9358.14  | -2.781  | -1397.71 | 3102.05 | 46.12    | -90083.1 | 1.9466  | 36.75   | 0   | -2.66   | -1.612   | 3.696  | -6.12   | -5.8736 | 1   |       |
| -0.15285 | 2893.33 | 1582.275 | 8637.52  | 2.16 | 8731.41  | -2.672  | -856.64  | 3290.55 | 19.734   | -89932.4 | 2.638   | 36.1    | 0   | -2.1247 | -1.624   | 0.5134 | -5.491  | -9.886  | 0   | 35.58 |
| -0.5688  | 2238.12 | 3454.32  | 11872.92 | 2.16 | 9824.61  | -3.378  | -1741.44 | 3842.15 | 44.452   | -89974.6 | 0.7786  | 4.32212 | 0   | -1.475  | -1.421   | 1.4014 | -5.0624 | -6.0488 | 0   |       |
| -0.4389  | 2850.23 | 1187.82  | 5144.92  | 2.16 | 9208.809 | -1.7223 | -1646.64 | 3333.45 | 28.672   | -90119.1 | 2.2766  | 11.418  | 0   | -1.4007 | -3.027   | 0.5788 | -5.2098 | -6.4232 | 1   |       |
| -0.855   | 2045.22 | 3290.82  | 8124.32  | 5.16 | 10101.21 | -2.3146 | -1426.73 | 2920.35 | 37.31006 | -89715   | 1.1158  | 18.74   | 1   | -1.5898 | -0.8308  | 2.1432 | -4.87   | -6.596  | 0   |       |
| -5.94    | 2756.58 | 4072.62  | 8109.12  | 5.16 | 2511.21  | -2.4887 | -1329.1  | 3132.92 | 27.024   | -90130.1 | 3.882   | 10.839  | 1   | -3.063  | -0.42001 | 3.752  | -4.7924 | -13.178 | 0   | 42.15 |
| -0.06878 | 2677.82 | 1897.02  | 8001.1   | 2.16 | 9466.74  | -1.6358 | -1414.13 | 3002.95 | 41.038   | -89977.9 | 1.4156  | 9.473   | 1   | -5.395  | -1.872   | 1.9534 | -7.01   | -9.848  | 0   |       |
| -0.27769 | 2711.61 | 1611.61  | 8781.12  | 5.16 | 8854.41  | -2.227  | -1333.66 | 3334.1  | 46.066   | -89969.3 | 1.0388  | 14.114  | 1   | -1.4377 | -1.1173  | 1.6686 | -5.4186 | -6.6128 | 1   | 37.44 |

For the sample the sake of machine learning predictions, a snippet of the above dataset was selected split for training and test before applying decision trees on the same set.

| population | amount \$ '000 | profitable |
|------------|----------------|------------|
| 2577       | 2183.52        | TRUE       |
| 1200       | 1654.74        | FALSE      |
| 2769       | 2661.42        | FALSE      |
| 3162       | 2199.12        | TRUE       |
| 2707       | 2083.62        | TRUE       |
| 2717       | 2248.02        | TRUE       |
| 2792       | 1338.12        | FALSE      |
| 1891       | 3179.82        | TRUE       |
| 2477       | 2659.92        | FALSE      |
| 2914       | 1502.7         | TRUE       |
| 2706       | 1513.47        | TRUE       |
| 2648       | 2289.72        | TRUE       |
| 691        | 1786.98        | TRUE       |
| 2893       | 1582.275       | TRUE       |
| 2238       | 3454.32        | TRUE       |
| 2850       | 1187.82        | FALSE      |
| 2045       | 3290.82        | TRUE       |
| 2757       | 4072.62        | TRUE       |
| 2678       | 1897.02        | TRUE       |

Further the dataset contains a total of 1000 instances and three columns of variable type character, float and Boolean.

## Analysis and presentations

Load and preview of the dataset on Jupyter IDE appears as below:

```
In [3]: import pandas as pd
sales_data_project = pd.read_csv('store_sales.csv')
sales_data_project
```

Out[3]:

|     | F1        | F2        | F3      | F4      | F5   | F6       | F7       | F8       | F9      | F10    | ... | F13     | F14 | F15     | F16     | F17    | F18     | F19      | F20 |
|-----|-----------|-----------|---------|---------|------|----------|----------|----------|---------|--------|-----|---------|-----|---------|---------|--------|---------|----------|-----|
| 0   | -0.123210 | 2577.3200 | 2183.52 | 7747.32 | 2.16 | 9693.21  | -2.24980 | -1381.98 | 3033.55 | 53.796 | ... | 6.7860  | 1   | -3.9290 | -0.8095 | 3.5800 | -5.9384 | -9.3420  |     |
| 1   | -2.064600 | 1200.1200 | 1654.74 | 6027.72 | 2.16 | 9391.26  | -2.74200 | -1336.20 | 2081.35 | 33.514 | ... | 5.5240  | 0   | -1.6609 | -2.3020 | 3.8880 | -6.5640 | -6.3806  |     |
| 2   | -0.083748 | 2769.3500 | 2661.42 | 7798.66 | 2.16 | 9296.28  | -3.21300 | -1531.16 | 3397.15 | 39.430 | ... | 10.0870 | 0   | -2.5100 | -2.8190 | 2.1606 | -5.2340 | -9.6780  |     |
| 3   | -2.808300 | 3161.5200 | 2199.12 | 8369.32 | 5.16 | 7990.71  | -2.27030 | -1397.03 | 3289.99 | 42.012 | ... | 15.5400 | 1   | -2.2555 | -0.9770 | 8.2160 | -5.8166 | -6.1512  |     |
| 4   | -0.445800 | 2707.4200 | 2083.62 | 9054.72 | 5.16 | 8691.81  | -3.73100 | -804.04  | 1903.35 | 43.172 | ... | 4.5188  | 0   | -1.7843 | -1.3690 | 0.7562 | -6.4060 | -9.5320  |     |
| ... | ...       | ...       | ...     | ...     | ...  | ...      | ...      | ...      | ...     | ...    | ... | ...     | ... | ...     | ...     | ...    | ...     | ...      | ... |
| 995 | -0.637200 | 2575.9200 | 1493.97 | 7726.38 | 2.16 | 8822.31  | -2.27120 | -1052.44 | 2790.55 | 56.832 | ... | 5.8160  | 0   | -2.0949 | -1.9290 | 1.3156 | -6.3420 | -6.9018  |     |
| 996 | -0.062512 | 2713.8200 | 1484.64 | 7529.12 | 2.16 | 9395.07  | -2.43980 | -1279.56 | 3143.73 | 49.938 | ... | 13.6120 | 1   | -1.6278 | -0.5306 | 0.7642 | -6.1828 | -8.1960  |     |
| 997 | -0.458100 | 2827.8675 | 1539.30 | 7463.92 | 5.16 | 9285.81  | -2.71500 | -1503.12 | 3497.55 | 20.112 | ... | 5.0459  | 0   | -1.8705 | -2.7190 | 2.2880 | -5.9302 | -8.0580  |     |
| 998 | -0.323700 | 2681.1200 | 2399.52 | 8268.92 | 2.16 | 10062.51 | -2.79400 | -1387.82 | 3001.85 | 32.880 | ... | 7.8240  | 1   | -3.1360 | -0.9401 | 4.0980 | -5.7562 | -12.4580 |     |
| 999 | -2.061000 | 4034.1200 | 1041.12 | 9405.32 | 5.16 | 8583.51  | -1.64566 | -1311.16 | 3986.95 | 29.592 | ... | 7.8380  | 0   | -1.8624 | -1.8080 | 1.5582 | -4.6114 | -6.1788  |     |

1000 rows x 22 columns

Further, the selected dataset was loaded and using pandas previewed the dataset as below showing the instances as 1000 with the expected 3 columns of variable names.

```
In [8]: import pandas as pd
#import sklearn.
sales_data_project = pd.read_csv('store_v2.csv')
sales_data_project.shape
```

Out[8]: (1000, 3)

In [ ]:



```
In [5]: import pandas as pd
sales_data_project = pd.read_csv('store_v2.csv')
sales_data_project
```

Out[5]:

|     | population | amount \$ '000 | profitable |
|-----|------------|----------------|------------|
| 0   | 2577       | 2183.52        | True       |
| 1   | 1200       | 1654.74        | False      |
| 2   | 2769       | 2661.42        | False      |
| 3   | 3162       | 2199.12        | True       |
| 4   | 2707       | 2083.62        | True       |
| ... | ...        | ...            | ...        |
| 995 | 2576       | 1493.97        | False      |
| 996 | 2714       | 1484.64        | False      |
| 997 | 2828       | 1539.30        | False      |
| 998 | 2681       | 2399.52        | True       |
| 999 | 4034       | 1041.12        | True       |

1000 rows × 3 columns

```
In [ ]:
```

A quick summary of the dates was also established by getting the total count the min value, the maximum value, the mean, standard deviation and the 25<sup>th</sup> and 75<sup>th</sup> percentile values of the dataset respectively

```
In [6]: import pandas as pd
sales_data_project = pd.read_csv('store_v2.csv')
sales_data_project.describe()
```

Out[6]:

|       | population   | amount \$ '000 |
|-------|--------------|----------------|
| count | 1000.000000  | 1000.000000    |
| mean  | 2520.915000  | 2120.448916    |
| std   | 747.502021   | 1587.054314    |
| min   | -2460.000000 | -7321.080000   |
| 25%   | 2440.000000  | 1505.370000    |
| 50%   | 2709.500000  | 1762.455000    |
| 75%   | 2836.000000  | 2227.920000    |
| max   | 6563.000000  | 18814.920000   |

## Prediction with Decision trees

The selected dataset was then supplied into the sklearn algorithm system to try to establish the expected values from the inputs. Given the trends established from the machine learns in the 1000 instances of train data. The researcher supplied two case scenarios of data containing the population of the area and the possible sales revenue of the location and tried to predict what would be the possible outcome of the supplied datasets and the results were as follows:

| Supplied data | Predicted output |
|---------------|------------------|
| 100,5000      | TRUE             |
| 3500, 600     | FALSE            |

```
In [11]: import pandas as pd
from sklearn.tree import DecisionTreeClassifier
sales_data_project = pd.read_csv('store_v2.csv')
X = sales_data_project.drop(columns=['profitable'])
Y = sales_data_project['profitable']
model = DecisionTreeClassifier()
model.fit(X, Y)
predictions = model.predict([ [100,5000],[3500, 600] ])
predictions
```

```
C:\Users\2395648\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.9_qbz5n2kfra8p0\LocalCache\local-packages\Python39\site-packages\sklearn\base.py:445: UserWarning: X does not have valid feature names, but DecisionTreeClassifier was fitted with feature names
  warnings.warn(
```

```
Out[11]: array([ True, False])
```

```
In [ ]:
```

## Interpretation and conclusion

From this predictive analysis, it can be observed that there is no direct relationship between the population and the sales revenues of the particular locations. The two variables are inversely related and none depends on the other. Actually if a correlation coefficient was done on this data, it would produce negative value of say -0.60, to just show that a change in the population or the revenue collections of the shop location does not in any way affect the performance of the shop in that particular location. Perhaps other factors such as market forces, legal implications and price demand variations of goods and products could be studied or looked at further in the next research.



