Predictive Data Analytics Martin Byrne Assignment 1

**Introduction**

The dataset being used for Regression tests is the TMDB 5000 Movie Dataset from Kaggle (The Movie Database (TMDb, 2017). There are 2 files in the dataset, however only the tmdb\_5000\_movies.csv file is used.

On to classification and a bank churn dataset (Mehmet A, 2020) is used which shows how customers use and depart a bank which gives the opportunity to see if there are any specific groups that this is happening to as well as finding out other factors in why they would stop using this bank.

Both datasets have sufficient information within them to allow for proper exploration and testing and have multiple columns and types available.

**Exploration**

The first section that we will look at is the exploration section and it is self-explanatory as it is where we will examine our data set to see what it contains and whether any operations will be needed to clean up the set so that it is more usable. We will only show the movies dataset exploration

The first step in the Jupyter notebook is to load the libraries that provide functionality, the screenshot shows the library import process.

Graphical user interface, text, application

Description automatically generated

The next step then is to import our dataset so that we may examine it and check out the first few lines using the head() method.

Graphical user interface, text

Description automatically generated with medium confidence

After doing our imports, we can check some basic information on the dataset using the info() method which allows us to see what types of data are being used.

Table

Description automatically generated

After this we can see some basic statistical information on the dataset using the describe() method

Table

Description automatically generated

This is sufficient as far as general exploring is concerned, we will need to do further exploration and analysis of the dataset but that will be more specific and belongs in each of the test cases.

**Regression**

The first thing that we must first do is explain what Regression is and why it is used. Regression is a technique that is used to analyse relationships between variables in a dataset and any outcomes that they lead to. These relationships should allow us to predict further outcomes so that we can make educated decisions on things like business investment, house prices, potential profits, and a lot more.

The first type of regression being used is basic linear regression and it will be used to check the relationship between a movie budget and how much it has earned in revenue to see if there is a correlation between budget size and revenue which I am of the opinion it will.

Firstly, make a copy of the dataset to operate on and then select the necessary columns as we do not need every column in the dataset. We only need 2 columns from the dataset, and we get these using the iloc() method with looks at the index location starting from 0, we can check them if necessary, using print().

Graphical user interface, text, application, email

Description automatically generated

We can then begin the process of splitting our data into test and training data, the test size standard is about 20% of the dataset and we can import the libraries for performing the Linear Regression.

Graphical user interface, text, application

Description automatically generated

We can then look at predicting the results and showing them on graphs, to do this we will look at the scatter plots for both test and training sets.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Then we can look at some things like the intercepting point, but the main point of interest is how precise the algorithm is and in this case the answer is not very as can be seen below.

Graphical user interface, text

Description automatically generated

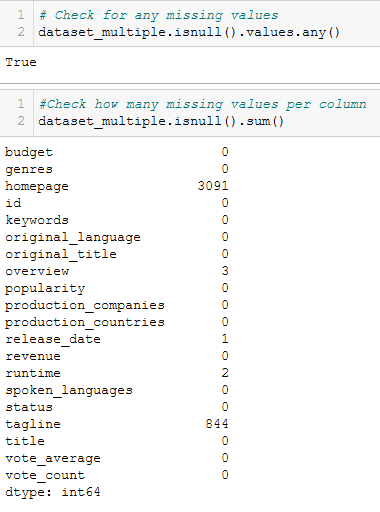
The simple linear model has a precision of about 53% which means that we would only get good revenue simply by having a big budget in slightly more than half the cases which also means that my own opinion would only be supported some of the time. This measurement is also known as R-Squared or coefficient of determination and shows the goodness of fit in the Linear Regression.

The second type of regression being looked at is Polynomial Regression, the reasoning behind this choice is that it will allow for examining whether some of the other numerical columns in the dataset such as runtime, how highly the movie is rated and do these factors have a relationship with each other as well as checking if the model performs better than the simple linear choice. Again, make a copy of the dataset and select the columns to be used in the modelling.

Graphical user interface, text, application

Description automatically generated

We must check if there are any missing or null values which need to be cleaned up. There are in the dataset, but they do not affect the columns that we are using in this case.



After this then we split our dataset as usual into test and train sets. Then scaling is done to avoid any issues with different scales and now is when the library needed for the Polynomial testing is performed.

Text, letter

Description automatically generated

After this then we will predict test results and evaluate the model performance.

Graphical user interface, text, application

Description automatically generated

This shows that after using several of the numeric columns that model performance is 51% which is not too good and worse than simply using Linear regression.

**Classification**

The first thing that we must do is to explain that classification differs from regression as regression uses continuous variables such as house prices whereas classification deals with categories such as whether a house is a cottage or town house, there is no continuous variable here, either a house is a cottage, or it is not.

The first classification model being looked at is Support Vector Machine, here we will use the bank churn dataset. It will be the case here and in most dataset testing and modelling that there are a lot of repeated steps.

Import and print dataset

Graphical user interface, text, application, email

Description automatically generated

After this then we will select the columns for our testing, the final column will be the dependent variable as it indicates whether a customer has exited the bank or not.

Text

Description automatically generated

Then we can perform the splitting of data into train and test sets.

Text

Description automatically generated

After this we will perform a scaling operation which will bring data to a common scale size, and which avoid issues which would otherwise arise.

Text, letter

Description automatically generated

We can then begin training the SVM model using the SVC classification library which is needed as we are checking a categorical variable in whether someone has stopped using a bank or not

Text

Description automatically generated with medium confidence

After this we can then predict test results and check our model accuracy score

Text, letter

Description automatically generated

In this case the accuracy score is relatively good at 79.64 percent which is good when trying to predict which customer will leave or stay which is very important as it gives a chance to try and stop customers who would leave by offering them better services.

We can also predict whether a person is going to stop being a customer when checking certain factors such as balance, salary etc. The screenshot shows that a customer is likely to f they are 10 years with the bank with a balance of 75 thousand, use 1 product, does not have a card and a salary of 90k. A return value of 1 indicates that the person will exit.

Text, letter

Description automatically generated

The second classification model being looked at is the Decision Tree classifier using the bank customer churn. The initial steps are the same as before so there is no need for screenshots other than when we are checking the test predictions score. We will also check our prediction to see if we can detect whether a customer will leave under certain circumstances. Many settings being used for SVM testing will be used by the Decision Tree classifier, this means that both classifiers are being run on the same factors and any difference will be down to the technique.

We start by using a different classifier.

Text

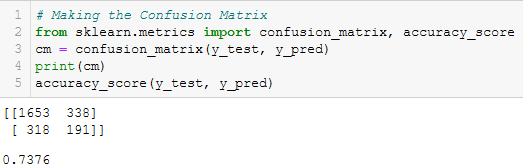
Description automatically generated

We can then run through predicting the test results.

A picture containing logo

Description automatically generated

Finally, then we can see how accurate the model is.



At over 73% this model is not bad in terms of accuracy but is slightly less than when using the SVM model at 79%.

One final screenshot is where we test again similar factors to see if a person is likely to leave the bank and again the answer is yes.

Text, letter

Description automatically generated

**Conclusion**

Some conclusions that can be drawn for the experiments above, starting first with the classification models which used the customer churn dataset. First is that the SVM technique outperforms the Decision tree technique in terms of model accuracy scoring 79.6 to 73.7. This is not a huge margin but may indicate that SVM is the better choice. The second is that there seems high level of churn from this dataset which means that the bank involved would have a lot of work to do to stop it occurring.

The regression techniques that were used which were used on the movies database, the first technique that was used was a basic linear technique where we checked for a correlation between the budget for a movie and how much revenue it earned. I had a personal feeling prior to calculations that this would be a high score but was proven wrong with only 53.2% accuracy. The Polynomial Regression also performed relatively poorly at 51% accuracy although it did test several factors rather than a single factor.

**References**

The Movie Database (TMDb (2017). *TMDB 5000 Movie Dataset*. [online] Kaggle.com. Available at: https://www.kaggle.com/tmdb/tmdb-movie-metadata?select=tmdb\_5000\_movies.csv [Accessed 29 Jul. 2021].

Mehmet A (2020). *Churn for Bank Customers*. [online] Kaggle.com. Available at: https://www.kaggle.com/mathchi/churn-for-bank-customers# [Accessed 31 Jul. 2021].

Select ONE Dataset

2 Regression Models

2 Classification Models

Justify and rationalize the choice of your models

Explanation of each Models analysis process

Prediction/Classification with the models

Conclusions

Total report Min 1500 words +/- 10%. Word count does not include code, comments, titles, or

references