|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Leeds University**  **Business School** | | A close-up of a sign  Description automatically generated | | | | | | | | |
| **Assessed Coursework Coversheet** | | | | | | | | | | |
| For use with *individual* assessed work | | | | | | | | | | |
| Student ID Number: | 2 | | 0 | 1 | 7 | 7 | 1 | 0 | 6 | 7 |
| Module Code: | LUBS5309M | | | | | | | | | |
| Module Title: | Forecasting and Advanced Business Analytics | | | | | | | | | |
| Module Leader: | Panagiotis Stamolampros | | | | | | | | | |
| Declared Word Count: | 2936 | | | | | | | | | |
| Please Note:  Your declared word count must be accurate, and should not mislead. Making a fraudulent statement concerning the work submitted for assessment could be considered academic malpractice and investigated as such.  If the amount of work submitted is higher than that specified by the word limit or that declared on your word count, this may be reflected in the mark awarded and noted through individual feedback given to you.  It is not acceptable to present matters of substance, which should be included in the main body of the text, in the appendices (“appendix abuse”).  It is not acceptable to attempt to hide words in graphs and diagrams; only text which is strictly necessary should be included in graphs and diagrams. | | | | | | | | | | |
| By submitting an assignment you confirm you have read and understood the University of Leeds [**Declaration of Academic Integrity**](http://www.leeds.ac.uk/secretariat/documents/academic_integrity.pdf) ( <http://www.leeds.ac.uk/secretariat/documents/academic_integrity.pdf>). | | | | | | | | | | |

**REPORT 1:**

*1.1 Data Handling*

The data provided to us is of Personal Consumption Expenditure or CPE, the data is initially loaded as a csv file in R studios, upon inspection for missing values, we observe 43 missing data entries which requires attention to handle them with appropriate method.

The missing values are imputed using the imputeTS library, the function used for imputation is na\_ma, this function replaces the missing values by the moving average values, the type of moving average is of the type ’weighted’ which gives higher weightage to recent events and the period of moving average is set as 12.

The above step ensured appropriate management of missing data, which resulted in a data set which has no missing data and can be considered clean.

Upon the cleaning of the data, the data is transformed into a time series object by using the ‘ts’ function, time series objects are data which are indexed by time (plotted against time)

*1.2 Forecasting models*

* Time- Series Generation and Plotting: The first step for forecasting required us to generate and visualise the time series for general understanding of the trend

A graph showing the growth of a company

Description automatically generated

*Fig 1.1 Time-Series of PCE*

The overall trend of the series is upwards, for a smoother understanding of the trend we shall observe the time series as a moving average depiction, the time-period of the moving average has been set to 21 observations, the resulted visual depicts a much smoother upward trend of the variable and gives us a clearer picture of the trend by overcoming any randomness and outliers.

A graph showing the growth of a company

Description automatically generated

*Fig 1.2 Moving Average 21 PCE*

* Decomposition of Time-Series: The Time series is composed of different aspects as we know (Trend, Seasonality, cyclical, residual/random), the decomposition will give us the insights into weather there exists any seasonal or cyclical components and if they are additive or multiplicative as this information will be essential for forecasting models.

A graph of a graph showing the same number of different types of numbers

Description automatically generated with medium confidence

*Fig 1.3 Decomposed Time Series*

We can observe a few bits of randomness in the time series which shall be delt later during forecasting.

Since the data is seasonally adjusted, we will chose to ignore the seasonal component.

* Seasonal Naive Model: Seasonal Naive is a simple forecasting technique, it accommodates a linear trend along with a seasonality component, where the future figure of seasonality is set as the old seasonal figure from the previous season in time, since the time series has both trend and seasonality, this naive type is best suited.

The data is split between train and test, in the ratio of 80 and 20 respectively, 20% of the data set comes out to be around 156, hence the training data is the entire data but the last 156 values from it, random sampling cannot be done as the test data should always be ahead in time than training data, this consideration is essential in terms of forecasting.

The plot of the forecast is as follows:

A graph of a line

Description automatically generated with medium confidence

*Fig 1.4 Seasonal Naive Forecast*

The above plot shows the forecast via the seasonal naive method , the forecast is done using the training data and the accuracy of it will be determined by comparing with the entire data of the PCE, the stagnant nature of the forecast is because of the naive model type as it just projects the previous seasonal value into future, the shaded area represents the uncertainty of the forecast.

The plot of the forecast with the original data is as follows:

A graph of a person

Description automatically generated

*Fig 1.5 Forecast vs Original Plot*

As we can observe from the above visual, the model was not the best in terms of capturing the actual flow of the data, suggesting a not so effective model type.

The accuracy of the model can be seen as below:



As it can be observed considerable errors can be witnessed with the test data, hence it shows further evidence of the inability of the model to forecast with efficiency.

* Holt Model: The Holt model is very versatile and is a model which accommodates the trend component, it is an extension of simple exponential smoothing, which accounts for trend component also and not just the levels component.

The plot of the forecast is as follows:

A graph showing the growth of a number of people

Description automatically generated

*Fig 1.6: Holt Forecast*

The uncertainty degree is observed to be more, but the forecast carries on the upward trend, the uncertainty funnel is wider due the incorporation of the seasonal component of the time series.

The comparison of the forecast with the original data can be seen as follows:

A graph showing a curve

Description automatically generated

*Fig 1.7: Plot of Holt forecast vs Original Data*

The forecast does a good job in accommodating the future values within the uncertainty funnel of the model, the model can be used for risk aversive events, the mean line of forecast approximately coincides with the original data for some time before the randomness of the original data takes place (mostly due to COVID 19).

The accuracy of the model can be stated as:



*Fig 1.8: Accuracy Holt Winter*

The accuracy shows decent performance, in the training set suggesting the model has fit well, however the errors in the test set are more in magnitude, suggesting that the model might not be able to generalise all the forecast values well.

* ARIMA: ARIMA stands for Auto Regressive Integrated Moving Average, it utilises regression of the previous data points to project the future values, the moving average component handles the error component of the time series, for ARIMA model the stationarity of the series is a must, hence differencing technique is used to stabilize the mean of the time series, this is the Integrated component which determines how many times the time series is differenced.

One-time differenced time series of PCE data can be seen below:

A graph of a person

Description automatically generated with medium confidence

*Fig 1.9: Differenced PCE time series*

We can see that the differencing has caused the mean of the time series to become constant, hence converting it to a stationary time series. In ARIMA we need to input the vectors of AR, I & MA, in the format as a vector (AR, I,MA) , to find the most appropriate order we shall run an auto.arima function which will give us the best possible order.

A close up of a computer screen

Description automatically generated

The above figure is the output of the auto Arima model, hence the order of (3,2,2) is best suited for our case, we then substitute the order in the actual ARIMA model and run the forecast.

A graph with a line

Description automatically generated

*Fig 1.10 ARIMA Forecast*

The above figure shows the forecast ran by the ARIMA model with orders (3,2,2), the shaded region is the uncertainty funnel, the model is used to forecast the next 156 values after the training set.

The below plot shows the plot of forecast with the original data:

A graph with a line going up

Description automatically generated

*Fig 1.11 Forecast vs Original ARIMA*

The tolerance of the forecast accommodates most of the original data, other than the sudden growth in the PCE, which goes beyond the uncertainty funnel of the forecast.

A screenshot of a graph

Description automatically generated

*Fig 1.12: Residuals for ARIMA*

The accuracy of the model is as follows:

A screenshot of a computer

Description automatically generated

The accuracy suggests the model has been trained well, there are few errors which can be noticed in the test set.

*2.0 October 2024 Values Forecast:*

The best model as per the accuracy comes out to be the holt model, hence we shall forecast the October 2024 values of PCE using this model, we shall use the entire PCE data in this forecast to maximise the accuracy of October 2024 forecast, upon running the model we get the following output:





The forecasted value for October is **19566.92** and the uncertainty funnel can be seen under Lo 80, Hi 80, Lo 95 and Hi 95.

*3.0 One-step Ahead Rolling Forecasting*

This model continuously updates data as the new data comes along, it generates forecast for the next figure while simultaneously updating the data, this step takes place consequently till the forecast is produced, we say ‘rolling’ because the origin of forecast keeps rolling ahead in time.( Clark, T.E. and McCracken, M.W., 2009)

Following are the accuracies of the models using this method:

1. Seasonal Naive



1. Holt



1. ARIMA



From the above the best model, given the factor for selection as lowest errors, is Seasonal Naive as it has the least ME and RMSE along with the least MAE.

* Conclusion:

We Conclude by stating that the best model is the Holt model and has given the best forecast, the original data was well within the forecasted funnel, and the errors for the model was least compared with the rest.

**The forecasted values by this model for October 2024 stands as 19566.92.**

**REPORT 2:**

* 1. *Introduction:*

Customer feedback is given utmost priority in the world of business as it can help overcome the shortcomings of the business.

This report aims to analyse such reviews of customers of a hotel, the report will identify the key factors that influence the stay o the customers, and this shall be done using topic modelling techniques.

* 1. *Data Description and Handling:*

The data give to us is the reviews and ratings of those reviews for a hotel, the initial step is to check for any missing values in the data set, upon inspection, no missing values were observed in either of the columns, the next step involved the sampling of the data, we sample 2000 reviews out of the data using the ‘sample\_n’ function (seed was set as the last three digits of ID).

After the sampling we had to split the data into positive reviews and negative reviews, the criteria used for the splitting was based on the review score achieved and it was as follows:

* Positive Data: If the review score was 4 or more than 4, we consider the reviews positive.
* Negative Data: If the review score was 2 or less than 2, we consider the reviews negative.

The review with ratings as 3 were considered neutral/moderate and hence were not part of the analysis.

* 1. *Data Pre-processing:*

Before the topic modelling, we need to ensure the text in the data is processed till a certain degree and ensure that the data is in a proper format for textual analysis, we undergo the following steps for the processing of the text:

* Tokenization: Tokenization breaks down the text into individual tokens, these are the simplest text units, which makes comprehension easy.



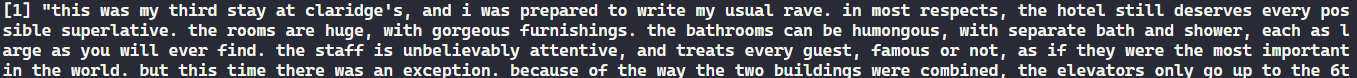
A group of white text on a black background

Description automatically generated

*Fig 1.1: Untokenized Vs Tokenized Text*

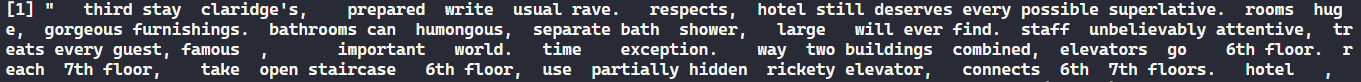
* Language Filtering: We only keep the reviews given in English language and discard the rest from the data.
* Corpus Formation: The next step involved the corpus formation of each negative and positive data using the ‘Corp’ function, corpus is a large collection of texts in a structured format.

* Conversion To lower case: This section uses the prompt ‘tolower’ in the ‘tm\_map’ function to transform the text into lower case, the complete text is reduced to lowercase format.



*Fig 1.2 Lower Case Transformed text.*

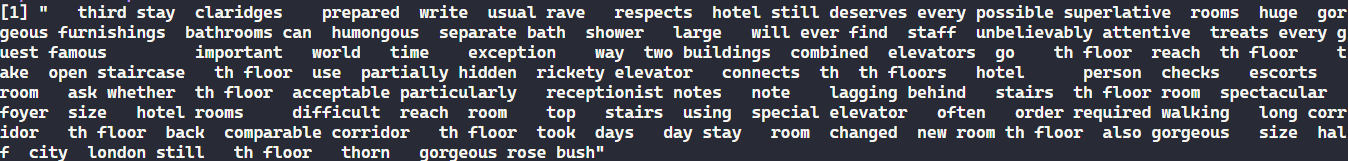
* Removal of Stop words: Stop words are words which occur often in text but do not hold any analytical knowledge, we remove stop words using the prompt ‘removeWords’ in the ‘tm\_map’ function.



*Fig1.3 After Removal of Stop Words.*

* Removal of Numbers: It is necessary to remove any numbers in the text as they can cause complication in text analysis later, this was done using the prompt ‘removeNumbers’ of the ‘tm\_map’ function, this enables removal of numbers from both corpuses the negative & positive.
* Removal of Punctuations: Punctuations hold no real insight about the texts and should be removed for more efficient textual analysis, the ‘removePunctuation’ prompt is used, which removes all the punctuations from the corpus.

After following the above steps, we get the cleaned and processed data which now can be used for the generation of Document Term Matrix.



*Fig 1.4 Processed Text*

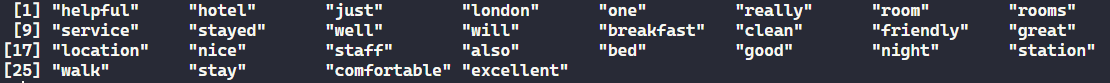
* 1. *DTM Formation:*

The Document-Term Matrix is nothing but a mathematical depiction of the text in a corpus in which rows are corresponded to each document within the corpus, and the columns represent the terms/words, the frequency of the terms is recognized due to the virtue of tokenization, we shall use this DTM for the formation of a word cloud , to get a visual depiction of the most frequent words in both negative and positive corpuses.( Yang, J. and Watada, J., 2011)

The function ‘DocumnetTermMatrix’ is used for the formation of the matrix for both the negative and positive DTMs, since the matrix is a sparse matrix (most of the entries are blank), we remove the sparsity by using the ‘removeSparesTerm’ function.

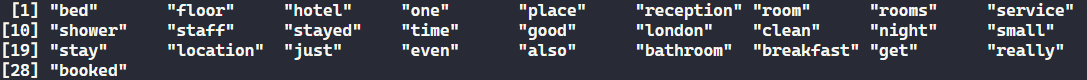
The DTM is used to find the most frequent words inside the corpus, which are as follows:

* Positive Corpus: We set the frequency to a count of 350 (words which occur 350 or more than 350 times in the corpus will be visible).



*Fig 1.5 Frequent terms in Positive Corpus.*

* Negative Corpus: We set the frequency as 60, frequency is set by trial.



*Fig 1.6 Frequent terms in Negative Corpus.*

* Word Cloud Formation: following are the word clouds for both the DTMs and can help us visualize the terms which occur more frequently in the Corpuses.

A close up of words

Description automatically generated

*Fig 1.7 Positive Word Cloud*

We can see that the terms like room, Hotel, breakfast, service, occur frequently in the positive corpus, suggesting that these are the terms which often gives a positive experience to the customers during their stay.

A close-up of a hotel

Description automatically generated

*Fig 1.8 Negative Word Cloud*

We can notice that terms like hotel, room, staff, shower, breakfast are most frequently occurred in negative reviews of the client, suggesting that these sections might need some improvement from the Hotel’s side.

*2.0 Topic Modelling:*

The process of topic modelling is carried out with LDA or Latent Dirichlet Allocation, it is a model based on probability, and is used to find topics which occur in documents, the process can be simply understood as categorization where LDA states that a document contains various broad topics and each word in the document can be related to one of the topics.( Ng, A.Y. and Jordan, M.I., 2003)

The graph uses three individual metrics: Griffiths2004, CaoJuan2009 & Arun 2010

From these metrics we need to find a satisfactory point in which the goal is to maximise Griffiths2004 metric, while minimizing the other two metrics.

We run a function called FindTopicsNumber and then plot the result to find the appropriate topic numbers.

* Positive Data Plot:

A graph of a graph

Description automatically generated with medium confidence

*Fig 2.1 Positive Data Plot LDA*

From the above plot we can infer that at the mark of 20 (number of topics), the Griffith metrics is at the highest point, and from the other to metrics which need to be minimized, both of them are at their lowest positions respectively at 20, hence this point becomes the most suitable point where the minimum and maximum coincide the best.

Hence, the **number of topics appropriate for the positive data is 20.**

* Negative Data Plot:

A graph with lines and dots

Description automatically generated

*Fig 2.2: Negative data Plot LDA*

From the above graph we can note that the point where the maximizing and the minimizing metrics are the closest is that of 14, as Griffith is at its maximum while the minimizing metrics are also sufficiently low, with CaoJuan metric at its 2nd lowest point.

Hence, the **number of appropriate topics for Negative Data is 14.**

*2.1 Topic Labelling:*

In this section we visually look at the terms under neath the topics and determine manually what that topic should be labelled as, the label should be in correlation with the terms beneath the topic.

* Positive Topics Labelling:

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer program

Description automatically generated

*A screenshot of a computer program

Description automatically generated*

*Fig2.3 Terms under 20 topics of Positive data*

* Topic 1: **Staff & Service**
* Topic 2: **Room Facilities**
* Topic 3: **Travelling Accessibility**
* Topic 4: **Positive Recommendations**
* Topic 5: **Worth of Money**
* Topic 6: **Staff Interaction**
* Topic 7: **Night Life**
* Topic 8: **Comfort & Luxury**
* Topic 9: **Stay Experience**
* Topic 10: **Hotel Facilities**
* Topic 11: **Dinning**
* Topic 12: **Room Hygiene & Comfort**
* Topic 13: **Breakfast**
* Topic 14: **Check-in & out Experience**
* Topic 15: **Room Noise & Ambience**
* Topic 16: **Timely Experiences**
* Topic 17: **Room Space/Area**
* Topic 18: **Environment around Hotel**
* Topic 19: **General Positive Remarks**
* Topic 20: **Locations**
* Negative Topic Labelling:

A screen shot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

*Fig 2.4 terms under the 14 Topics of Negative Corpus.*

* Topic 1: **General Experience**
* Topic 2: **Building Structure**
* Topic 3: **Night Stay**
* Topic 4: **Front Desk Service**
* Topic 5: **Locality**
* Topic 6: **Room types**
* Topic 7: **Room Quality**
* Topic 8: **Hotel Food Experience**
* Topic 9: **Booking Experience**
* Topic 10: **Staff Interaction**
* Topic 11: **Room Bed & Facilities**
* Topic12: **Locations Experience**
* Topic 13: **Bathroom Facilities**
* Topic 14: **Breakfast experience**

*2.2 Topic* Probability: Now we shall find the most probable topics in the corpus for each positive and negative corpus, by this virtue we can then find the top three topics for both negative reviews and positive reviews.

The Probability matrix for the positive data is as follows:

A screen shot of a computer screen

Description automatically generated

A screen shot of a computer screen

Description automatically generated

*Fig 2.5 Probability Matrix for Positive*

From the above Figure we can observe that the most probable topics in the above documents are topic 9, 5 and 1, based on their mean probability calculations.

The Probability Matrix for the Negative data is as follows:

A screen shot of a screen

Description automatically generated

A screen shot of a computer screen

Description automatically generated

*Fig 2.6 Probability Matrix for Negative*

From the above figure we can infer that the topics 14, 2 & 9 are the most probable topics.

* 1. *Conclusion:*

From the above textual analysis, we have come to a conclusion as to what are the top three factors that affect the customers positively and what three factors effect the customers negatively.

* Factors Effecting Positively:

1. Stay experience.
2. Worth of Money
3. Staff & Service

* Factors Effecting Negatively:

1. Breakfast Experience
2. Building Structure
3. Booking experience

Hence the Hotel should try and improve the breakfast and the booking procedures and at the same time start renovation of building structures, and they should maintain the standards of the staff, and general stay experience while keeping it in the budget of the customer.

References:

Yang, J. and Watada, J., 2011, June. Decomposition of term-document matrix representation for clustering analysis. In *2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011)* (pp. 976-983). IEEE.

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), pp.993-1022.

Clark, T.E. and McCracken, M.W., 2009. Improving forecast accuracy by combining recursive and rolling forecasts. *International Economic Review*, *50*(2), pp.363-395.