**CLUSTERING TECHNIQUES (NLM3 TASK 1)– D213**

**Performance Assessment**

**Western Governors University**

**By: Christian LeBlanc**

***Part I: Research Question***

***Does revenue have any recurring seasonal patterns?***

A.  Describe the purpose of this data analysis by doing the following:

1.  Summarize **one** research question that is relevant to a real-world organizational situation captured in the selected data set and that you will answer using time series modeling techniques.

2.  Define the objectives or goals of the data analysis. Ensure your objectives or goals are reasonable within the scope of the scenario and are represented in the available data.

***Part II: Method Justification***

B.  Summarize the assumptions of a time series model including stationarity and autocorrelated data.

Time series models have assumptions that include stationarity and autocorrelated. The assumption of stationarity is the data follows a consistent behavior over time. The assumption of autocorrelated is that data points are correlated with data points from the past.

***Part III: Data Preparation***

C.  Summarize the data cleaning process by doing the following:

1.  Provide a line graph visualizing the realization of the time series.

In preparing the data I had to format the “Day” into “Date” as it is necessary to have a format such as year-month-date to be able to see seasonal trends instead of having a count of date that does not give information on the time of year. No gaps are present in the above graph. The length of the sequence is from January 1st 2022 to January 1st 2024 with a total of 731 days.

2.  Describe the time step formatting of the realization, including *any* gaps in measurement and the length of the sequence.

3.  Evaluate the stationarity of the time series.

Using pandas I imported the teleco time series csv file with Day being my index column. I used info and describe to look at the data’s characteristics and check for nulls. I changed the day column into Dates. Then I checked if the data has stationarity. It did not, so I used the differencing function. I checked for stationarity again and this time the results came back that it is stationary. I will export this clean data in a csv file as it is a requirement in C5. Then I will split the data into training and test sets and export each of these into a csv file as required.

4.  Explain the steps you used to prepare the data for analysis, including the training and test set split.

5.  Provide a copy of the cleaned data set.

***Part IV: Model Identification and Analysis***

D.  Analyze the time series data set by doing the following:

1.  Report the annotated findings with visualizations of your data analysis, including the following elements:

The above visualization of a subset of the data shows no obvious seasonality.

The spectral density graph above shows a lack of seasonality due to the spikes in it.

The ACF of the data set before differencing shows a slow decrease in lag, which is another indication of the data being non-stationary.

The data set after differencing display most of the values in the ACF and PACF in the blue shaded area which means they are not statistically significant.

•   the presence or lack of a seasonal component

•   trends

•   the autocorrelation function

•   the spectral density

•   the decomposed time series

•   confirmation of the lack of trends in the residuals of the decomposed series

2.  Identify an autoregressive integrated moving average (ARIMA) model that accounts for the observed trend and seasonality of the time series data.

3.  Perform a forecast using the derived ARIMA model identified in part D2.

4.  Provide the output and calculations of the analysis you performed.

5.  Provide the code used to support the implementation of the time series model.

***Part V: Data Summary and Implications***

E.  Summarize your findings and assumptions by doing the following:

The ARIMA model was the correct model to select due to no seasonality being present in the time series. Using Auto-ARIMA, the best fit model is 1,0,0 with no seasonality 0,0,0, and no periodicity 0. The forecast prediction interval is 1 day. Our TM data consists of 2 years of daily revenue. Consequently, the ARIMA model identifies correlations and seasonality to predict revenue daily. The forecast is set to 90 days. The two years of revenue data is good enough to forecast up to a year of future revenue. Setting it to 90 days allows the results to be more accurate than the year but long enough to give good insights as many companies do quarterly revenue calls. As pointed out previously I used Auto-Arima to find the best fit model. Using the root mean squared error, RMSE, which resulted with approximately 0.561. A number closer to 0 would be ideal.

I would recommend that the company would take these results for the trend of revenue for the next quarter and see how they could expand the company to cause revenue to increase.

1. Discuss the results of your data analysis, including the following points:

•   the selection of an ARIMA model

•   the prediction interval of the forecast

•   a justification of the forecast length

•   the model evaluation procedure and error metric

2.  Provide an annotated visualization of the forecast of the final model compared to the test set.

3.  Recommend a course of action based on your results.

***Part VI: Reporting***

F.  With the information from part E, create your report using an industry-relevant interactive development environment (e.g., an R Markdown document, a Jupyter Notebook). Include a PDF or HTML document of your executed notebook presentation.

G.  Cite the web sources you used to acquire third-party code to support the application.

H.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

I.  Demonstrate professional communication in the content and presentation of your submission.

**Part I: Research Question**

A.  Describe the purpose of this data analysis by doing the following:

1.  Summarize **one** research question that you will answer using neural network models and NLP techniques. Be sure the research question is relevant to a real-world organizational situation and sentiment analysis captured in your chosen data set(s).

A1

Using previous reviews on a product or service, can we predict if a user has a positive or negative opinion on it?

Note: If you choose to use more than one data set, you must concatenate them into one data set for parts II and III.

A2

The goal of this analysis is to see if the words a user uses can predict the sentiment they have for a product or services.

2.  Define the objectives or goals of the data analysis. Be sure the objectives or goals are reasonable within the scope of the research question and are represented in the available data.

A3

Recurrent Neural Networks are used for text classification. RNN can produce text classifications such as sequence labeling, speech, tagging, and more.

3.  Identify a type of neural network capable of performing a text classification task that can be trained to produce useful predictions on text sequences on the selected data set.

**Part II: Data Preparation**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import nltk

from nltk.corpus import stopwords

from nltk.tokenize import word\_tokenize

import re

import warnings

warnings.filterwarnings('ignore')

from sklearn.model\_selection import train\_test\_split

import tensorflow as tf

import keras

import tensorflow.keras

from keras.layers import Embedding

from tensorflow.keras.models import Sequential

from tensorflow.keras.preprocessing.text import Tokenizer

from tensorflow.keras.preprocessing.sequence import pad\_sequences

from tensorflow.keras.models import Sequential

from tensorflow.keras.layers import LSTM,Dense, Dropout, SpatialDropout1D

from tensorflow.keras.layers import Embedding

import statsmodels.api as sm

B.  Summarize the data cleaning process by doing the following:

1.  Perform exploratory data analysis on the chosen data set, and include an explanation of each of the following elements:

•   presence of unusual characters (e.g., emojis, non-English characters)

•   vocabulary size

•   proposed word embedding length

•   statistical justification for the chosen maximum sequence length

I saw that the longest review length of 1390, which made me investigate the median and 75th percentile to get a better feel for the data. The 75th percentile of 16 in comparison to the max shows there is outliers. I will use 16 as the maximum sequence length to hopefully gain some accuracy in my model.

2.  Describe the goals of the tokenization process, including any code generated and packages that are used to normalize text during the tokenization process.

The tokenization process assigns a number or “tokens” to each word. I will be using one hot from TensorFlow. Keras package to accomplish this.

3.  Explain the padding process used to standardize the length of sequences. Include the following in your explanation:

•   if the padding occurs before or after the text sequence

•   a screenshot of a single padded sequence

4.  Identify how many categories of sentiment will be used and an activation function for the final dense layer of the network.

You can set the labels for your data set to 0 and 1. 0 = Negative review and 1 = Positive review. In that case, you have two categories that could be used for sentiment analysis output in the neural network output layer (Analytics Vidhya, 2021).

I imported all the three files and concatenated them. I renamed the columns to give better meanings to them. Then I removed the unusual characters by making all lowercase, removing the punctuations, and replaced the special characters with spaces. I created a new column for the length of the reviews. Next, I got the vocabulary number, longest review, and the recommended embedding size. Then I tokenized the reviews and padded them to a max length of 16. Below I will split the data into 80/20 for the training and test sets.

5.  Explain the steps used to prepare the data for analysis, including the size of the training, validation, and test set split (based on the industry average).

6.  Provide a copy of the prepared data set.

**Part III: Network Architecture**

C.  Describe the type of network used by doing the following:

1.  Provide the output of the model summary of the function from TensorFlow.

The model has five layers. The first layer is embedding with 216 parameters. The second layer is Flatten. The third is Dropout. The fourth and fifth layers are Dense. The fourth has 2,064 parameters and fifth has 17.

2.  Discuss the number of layers, the type of layers, and the total number of parameters.

3.  Justify the choice of hyperparameters, including the following elements:

•   activation functions

Both the Sigmoid function and the Rectified Linear Activation (ReLu) were used in the model. The Sigmoid function was used for its binary classification. The ReLu is efficient activation function used for hidden layers.

•   number of nodes per layer

The first three layers have 128 nodes. The third has 16. The last layer has 1 node.

•   loss function

The loss function for the model is Binary crossentropy due to the binary responses of positive or negative reviews.

•   optimizer

The “adam” optimizer is the optimizer for the model. It was chosen due to its performance and efficiency.

•   stopping criteria

I will use a stopping criteria of patience of 2. This will be done to avoid overfitting of the training data.

•   evaluation metric

Using the “accuracy” metric to evaluate the model gives a Test accuracy of approximately 52.7%.

**Part IV: Model Evaluation**

D.  Evaluate the model training process and its relevant outcomes by doing the following:

Above I ran a model fit with 30 epochs without any stopping criteria and then with a stopping criteria of 2 patience. The stopping criteria helped efficiency by dropping the number of epochs needed.

1. Discuss the impact of using stopping criteria to include defining the number of epochs, including a screenshot showing the final training epoch.

The training accuracy being approximately 10% higher than the validation and test accuracies shows some overfitting is present. This could possibly be improved by things such as a simpler model or having access to more data to improve the model’s ability to generalize.

2.  Assess the fitness of the model and any actions taken to address overfitting.

3.  Provide visualizations of the model’s training process, including line graphs of the loss and chosen evaluation metric.

4.  Discuss the predictive accuracy of the trained network using the chosen evaluation metric from part D3.

The accuracy of the trained network after the 4th epoch is 70.9%. This is when the early stopping of patience equal to two stopped. The graph shows that the validation accuracy mirrors the rise and falls but it does it around 8-10% lower.

**Part V: Summary and Recommendations**

E.  Provide the code you used to save the trained network within the neural network.

F.  Discuss the functionality of your neural network, including the impact of the network architecture.  
The data given has 2,748 reviews. 80% of them were used to train the model and the other 20% were used to test the model. The Natural Language Processing (NLP) is used to do a sentiment analysis on words that would indicate positive or negative reviews.

G.  Recommend a course of action based on your results.

I would recommend gathering more data to give more information to the model in the training data in hopes that it would cause the model to score a better accuracy especially in the test accuracy. In gaining more reviews I would assume that it would cause things such as the number of vocabulary to grow which would call for many changes in the model.

**Part VI: Reporting**

H.  Show your neural network in an industry-relevant interactive development environment (e.g., a Jupyter Notebook). Include a PDF or HTML document of your executed notebook presentation.

I.  Denote specific web sources you used to acquire segments of third-party code that was used to support the application.

J.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

K.  Demonstrate professional communication in the content and presentation of your submission.