Final Project Programming For AI
Journal

Diego Lemos
20204787

For this project my team and I have chosen the topic of Time-series analysis and forecasting in Retail industry, the dataset used by me on my part of the project was the Online Retail dataset , this is a transactional dataset that includes every transaction made by a UK based, registered non-store internet retailer between Dezember 2010 and Dezember 2011.

I started the project with the Data Loading and Processing, for this first part I spent between 3 to 4 hours, beginning with the load of the dataset and then the exploration of its information, as the datatype of its features and the amount of the observation and features. After that, I checked the missing values and only 2 features had values missing, 'Description' and 'CustomerID', for the first one, 'Description', as it was more valuable for future analysis and the number of missing values wasn't so high compared to the size of the dataset, I decided to drop it. For the second, 'CustomerID', I decided to replace the missing values with '0' as the number of rows with missing values was very high, and dropping them could affect the later analysis.

There were also a few duplicated rows that were dropped as it looked like to be the same entrance of data replicated. While working on outliers, it was noted some negatives values in 'Quantity' and 'UnitPrice' features, so they were filtered from the data frame, after that it was took in consideration that those negative values could represent returns made by customers, thus it was saved on a specific data frame for possible future analysis.

After plotting the boxplot, it was detected outliers in 'Quantity' and 'UnitPrice' features. Thus, the interquartile range technique was applied to eliminate the outliers in those columns. This statistical technique consists of taking the difference between the third quartile and the first one from a group of data then it measures the central dispersion of the data, eliminating the outliers.

To reduce memory usage during the analysis performance the datatype of features such as 'InvoiceNO', 'StockCode', 'Description', 'Country', and 'CustomerID', were optimized. Also, to for a better application of the Time-Series analysis, temporal features were extracted from the 'InvoiceData' column.

Finally, After plotting the distribution of the 'Quantity', 'UnitPrice' and 'Revenue' features, I noticed that 'Quantity' and 'Revenue' contained a very skewed

distribution, as it is showed in the notebook's plot, the long tail in the right side of it. Skewed data can dominate a model training, thus to avoid such a thing, log transformation was applied on these features, reducing variances, making large outliers less influential, and putting data closer to a normal distribution that will be ideal for future application of forecasting technique.

The second part of my project was the Exploratory Data Analysis (EDA), here I spent at least 7 hours, I started by checking the aggregate monthly revenue for the year of 2011, and it is possible to note that the online sales were down in the first semester and grows in the second, towards the end of the year and Christmas time, what makes its more interesting is that in November of 2011 the sales were way higher then Dezember, probably because of events like Black Friday witch can make people to by Christmas presents earlier. After that I plotted the aggregate of revenue by weekday, and it was possible to see that Thursdays brings in the biggest revenues on online sales, followed by Tuesday. Sunday, on the other hand, generates the last amount of sales. This can suggests that weekdays patterns may have an impact on sales, maybe reflecting the purchasing habits of customers.

Following the analysis, I plotted the top 10 countries by Revenue and then after that I repeated the plot but after performing a normalization by calculating total revenue and transaction per country, so we could see that, In the first plot was clear that UK dominates the revenue, significantly overshadowing the other countries as already expected, due that the dataset is from the UK based and registered non-store online retail, but it is interesting to note that when it get normalized by the number of transactions per country, it is noted that countries such as Netherlands, Australia and Singapore generates higher revenue per transaction compared to the UK.

After countries, I made an analysis in top customers and its behaviour, trying to understand their tends and patterns, I noticed that the customer with ID 14646 contributes with the most revenue among the top 10 customers, indicating their significance to the business. The customer with ID 14911 is the one with the highest number of transactions among the top ones, it shows how engaged he did with the platform, the customer with ID 16446 has a significantly longer average time between purchases, around 102 days, the business could set a remainder or even think about re-engagement campaigns. I also can see that the customer ID 14646 has a specific kind of product that he always buys, being the 'Spaceboy Lunch Box" the most bought.

The following analysis was through the products, Through the products' revenue analysis I could see that the product 'Paper craft, litte brirdie' stands out as the top performer and it contributes more revenue than others. For the percentage

contribution to total revenue we can note that the top 3 products contribute collectively more than 4% of the toral revenue. And for the monthly revenue trends it is possible to note that 'Paper craft, litter birdie' has a consistent contribution across months. The last analysis was comparing the quarters revenue, the fourth quarter is the stronger one, with more sales, mostly because of the holidays, but even though December has the Christmas time, November, probably because sales events such as Black Friday is the one with the most revenue.

The last part of the project was the TIME-Series and Forecasting, I have spent another 4 hours and I started applying the ANOVA test, The ANOVA test is a statistical method used to compare the means of three or more group and determine if there is differences between them. As per the results of the test, the difference between the quarters are not statistically significant, the p-value is much greater then the common threshold of 0.05 or 5% significance level meaning that there is no strong evidence to reject the null hypothesis. The null hypothesis assumes that the means of groups are equal, since the p-value is high, the data doesn't provide sufficient evidence to claim that the revenue differs significantly between quarters.

After, I plotted a seasonal heatmap to visualize the seasonal revenue patterns over 13 month helping on identifying the peak and low periods. Thought this heatmap of the revenue by month, I could note the gradual buildup to the peak season, the steady growing of the revenue starting from august and September leading to November indicates a preparation for the holiday season.

Following, I, used the ARIMA and Prophet models for forecasting, Even though there is a limitation on the data as it only has 13 month to be analysed, I implemented those models for forecasting and the values in the result represents the best the model's best estimates given the constrains od the dataset. To confirm the performance of the models, a validation was implemented, using the evaluation metrics, Mean Absolute Percentage (MAPE) and Root Mean Squared Error (RMSE). MAPE works calculating the average percentage difference between the predicted and actual value as the RMSE measures the standard deviation of the prediction errors, giving more weight to larger errors. For the ARIMA model the MAPE was 33.50%, indicating that the average prediction error is about to one-third of the actual revenue and the RMSE was 311,649.88, reflecting the average deviation of the predictions from the actual values, and for the Prophet model the MAPE was 36.60% and RMSE was 329,113.33 showing that the ARIMA got a slightly better results.

In conclusion for this project, I explored the Online Retail Dataset, that was alight with my team topic, o discover trends in its revenue, customer behaviour and product performance. Also develop a predictive model to forecast revenue.

Since the dataset only covers 13 months, it was very challenging to identify long-term seasonal patterns, for accuracy findings, many forecasting models, such as ARIMA or Prophet, need more data. Missing values needed to be handled carefully, especially in customer IDS, although this was resolved, bias can occasionally be introduced through imputation, furthermore, although ARIMA and Prophet performed rather well, they had trouble forecasting direct seasonal rise, like those that occurred in November and December. The error measures made through MAPE and RMSE indicates that while the predictions were accurate, they could yet be improved.