**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Data Visualisation Techniques  Machine Learning for Business |
| **Assessment Title:** | Individual |
| **Lecturer Name:** | David McQuaid  Sam Weiss |
| **Student Full Name:** | Zygimantas Jakubauskas |
| **Student Number:** | Sba22342 |
| **Assessment Due Date:** | 26/May/2024 |
| **Date of Submission:** | 26/My/2024 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Analysis of online retail business**

**Table of contents**

Introduction …………………………………………………………………….3

* + 1. About the data ………………………………………………………………….3
    2. Exploratory data analysis……………………………………………….……....3
    3. Data preparation………………………………………………………….…..…4
    4. Content and Collaborative filtering ……………………………………………4
  1. Collaborative Filtering………………………………………………..…………...5
  2. Content Filtering …………………………………………………………….….…6
  3. Evaluation, justification and recommendations…………………………………....7
     1. Time series Analysis……………………………………………………………7

5.1 Augmented Dickey-Fuller test results………………………………………………7

5.2 Implementation of ARIMA model………………………………………………….8

5.3. One step ahead forecast ……………………………………………………….…..10

* + 1. Visualisations……………………………………………………………..……11
  1. Data preparation for visualisation………………………………………………….11
  2. Interactive dashboard………………………………………………………………11
     1. Conclusion …………………….……………………...……………………….12
     2. References and sources…………………………………………..……….……13

**Table of figures and tables**

Figure 1. Corelation Heatmap……………………………..……………………….………..3

Figure 2. User-User recommendation table tested with customer 235143……………..…...5

Figure 3. User-User recommendation table tested with customer 235357……………….…6

Figure 4. Item-Item recommendation table tested using different variables in Rows…….....6

Figure 5. Statistical result of Augmented Dickey-Fuller test…………………………….….8

Figure 6. Autocorrelation graph of ARIMA model………………………………………….9

Figure 7. Partial Autocorrelation graph of ARIMA model……………………….…………9

Figure 8. Diagnostics plots of ARIMA model…………………………………………..….10

Figure 9. One-step-ahead prediction plot…………………………………………………….11

**Introduction**

This project looks at online sales of Bicycle online retailer. Its sales data is used to find best recommendation systems for its customer, to perform time series analysis and to create interactive dashboard aimed at 18-35 year old customers.

1. **About the data**

The main dataset used for this project is "Sales.csv". Dataset is publicly available on <https://www.kaggle.com/datasets/sadiqshah/bike-sales-in-europe> and looks into online business selling bicycles and their accessories around the world. Dataset contains data about their customers – their age and gender, and date when the order was placed, information about the products, their name, category and sub-category; and the financial information – cost of the items, their selling price, profit made on each transaction and the revenue details.

1. **Exploratory Data Analysis**

Dataset is checked for duplicates and missing values. It contains no missing values but contains 1000 rows of duplicates that are removed during data preparation.

Original dataset contains 18 variables and 113036 observations. Values are object or integer type.

Categorical and Continuous variables plotted using histograms and scatterplots. Here is the correlation heatmap.

A screenshot of a blue grid

Description automatically generated

*Figure1. Correlation Heatmap*

Heatmap shows few heavy correlations. “Profit” and “Cost” show strong positive corelation of 0.9. Other financial metrics such as “Unit\_Cost”, “Unit\_Price” and “Revenue” corelate positively strongly between themselves.

1. **Data Preparation**

Data preparation starts with removing duplicate observations. There are 1000 of them.

Next columns are renamed to give more user-friendly names for better understanding. As dataset contains no missing values, time to move on to perform tasks, where data is further prepared to meet tasks requirements.

1. **Content and Collaborative Filtering**

Content and collaborative filtering are two popular systems of recommending items to the users, normally online.

Collaborative filtering is based on idea that if one user like some certain items, another user who likes same items, would buy other items liked by the first user. Collaborative filtering focuses on personal parameters, normally ratings or reviews customers give.

Content filtering is based on idea, that if customer likes the item with some certain parameters or attributes, he’d like other similar products too. Parameters may include same or similar category of the product.

To run content (item-item) and collaborative (user-user) filtering dataset needs crucial details that are missing from this particular dataset. Therefore, two additional columns are created: “InvoiceID” and “CustomerID”.

**InvoiceID**. InvoiceID is created using variables like “Date”, “Age”, “Gender”, “Country” and “State”. The logic behind it is that if all this parameters match, products were sold on 1 invoice. If parameters don’t match completely, invoices would be different, as customers are different buyers.

**CustomerID**. It’s impossible to know exactly which customers are returning. However, the best guess is to filter customers using variables ““Age”, “Gender”, “Country” and “State”. If all these parameters match there is a decent possibility, it’s a returning customer, although inaccuracies are expected.

Once new variables are created, is possible to build pivot table. Rows are indexed by CustomerID. Columns are different products. The values in the table are the sum of quantities purchased by each customer. Newly created table has 2162x130 shape, and not having too large number of customers, which could have be seen as inaccuracy by the way variable CustomerID was created, now helps to keep pivot table not to big, which in turn enables to make further calculations on a full table unrestricted by computers technical abilities.

Why quantities and not ratings? The original dataset doesn’t have ratings, but quantities have a reasonable indication of the products popularity. Popular products normally mean good products. The most popular products may not necessarily get highest ratings, as those ratings possibly would be given to the most expensive items, but popularity certainly indicates a very good quality/price balance, meaning that customers are happy with frequently bought items. Therefore, it makes sense to use quantities in cases where ratings aren’t available.

With pivot table created it’s time to move to recommendations.

* 1. **Collaborative Filtering**

First to analyse is collaborative or user-user filtering. New matrix, using cosine similarity, is created. **Cosine similarity is a mathematical metric used to measure the similarity between two vectors in a multi-dimensional space, particularly in high-dimensional spaces, by calculating the cosine of the angle between them.** (Miesle, 2023)

**With model now built it’s time to test and assess its performance. It’s tested with 2 customers. First customer is the one who spent most money, and second customer is the one who bought most unique items. Both these customers are pitted against their nearest most similar customers, which similarity is calculated by cosine similarity. These are the results:**

A table with text and images

Description automatically generated

*Figure2. User-User recommendation table tested with customer 235143*

A table with text and numbers

Description automatically generated with medium confidence

*Figure3.* *User-User recommendation table tested with customer 235357*

* 1. **Content Filtering**

For content filtering 2 different approaches were taken. One is based on CustomerId and another is based comparing the products and using their sub\_categories.

A table with text and numbers

Description automatically generated

*Figure4. Item-Item recommendation table tested using different variables in Rows*

Table clearly indicates that results can be completely different depending on the way pivot table is built. The first column shows the results when pivot table was created using rows filled with CustomerID. Although recommendations seem odd, they actually make sense. There is a chance that customer that bought Classic Vest, may want to Water Bottle, or the fifth recommendation Logo Cap.

Second approach using Sub\_Categories as rows, provides different set of results. The recommended items are very similar to the originally bought item. But these recommendations aren’t suitable for retail customers but may well suit wholesale customers.

* 1. **Evaluation, justification and recommendations**

Different methods analyses and different results are achieved. Which method to prioritise, heavily depends on the whole customer base. Customer segmentation is needed, especially important to separate wholesale and retail buyers as many big companies sell to both.

User – User method can work well most of the time, especially in retail, as customers who bought similar products are very likely to have similar taste and it makes sense to offer the products that aren’t in similar customer’s basket.

Item – Item method can produce completely different results, depending on the way pivot matrix is built but all the results can be successfully used by the seller if adapted correctly. In this analysis where as index CustomerID is used, recommended products should be recommended to odd casual buyers. Where as index Sub\_Category is used, items are very similar to the tested item and can’t be recommended to retail customers. This is a classical mistake, which is even constantly made by online retail giant Amazon, which offers to a customer that just bought a camera, to buy another camera. However this algorithm may well be used for wholesale customers, that want to have a lot of very similar items to expand their range.

1. **Time Series Analysis**

The dataset used in this project is suitable for time series analysis as contains date. Data is collected during 5-year period and gives an opportunity to look for trends and seasonality.

* 1. **Augmented Dickey-Fuller test results**

Augmented Dickey-Fuller test is an important statistical test to determine whether datasets data is stationary or not. To know stationarity is very important, as it enables picking the correct Box-Jenkins model to make forecast predictions. Fox example ARIMA model cannot forecast on non-stationary data. So ADF test if that dataset contains the unit root. The null hypothesis is that the dataset does contain the unit root and test results will determine if null hypothesis should be accepted or rejected.

After implementing ADF test following results were calculated:

A screenshot of a computer test results

Description automatically generated

*Figure5. Statistical results of Augmented Dickey-Fuller test*

The ADF Statistic is the first indicator of the presence of unit root, and as the ADF statistic value is more negative as more unlikely unit root exists.

p-value has to be below significance level of 0.05 for null-hypothesis to be rejected. And in this case p-value is 0.044, or less than 0.05.

The next two components show that 16 lags (delayed observations) and 1207 observations were used during the test.

Next three parameters are significance levels, and they need to be checked against ADF Statistic. If ADF Statistic is lower than critical value, null hypotheses can be rejected. In this case null-hypothesis can be rejected with 99% confidence.

Lastly Maximised Information Criterion indicates how good machine learning model will fit the dataset. As lower is AIC as better. While on its own, it’s difficult to judge AIC parameter, it becomes useful when comparing few combinations of dataset, few differently filtered datasets, to find the lowest AIC for better fitting model.

On the balance ADF test determined that dataset is stationary, and the Box-Jenkins model picked for further analysis is ARIMA.

* 1. **Implementation of ARIMA model**

Implementing of ARIMA model starts with plotting Autocorrelation and Partial Autocorrelation graphs.

A graph with red dots

Description automatically generated

*Figure6. Autocorrelation graph of ARIMA model*

A graph with red dots

Description automatically generated

*Figure7. Partial Autocorrelation graph of ARIMA model*

Here the X-axis represents time difference between time series, and it’s lagged values. Y-axis represents autocorrelation coefficients. The coloured area on the graph represents insignificant values. Graph does not show really significant values, although some value are somewhat significant. The value at lag 0 is always 1 since the series are perfectly corelated with itself at same time points. The value at lag 1 is 0.25 – somewhat significant, therefore order used in model fitting is (1,0,0).

After model is fitted summary and diagnostics can be extracted.

A collage of graphs and diagrams

Description automatically generated

*Figure8. Diagnostics plots of ARIMA model*

What do diagnostics explain. Standardized residuals are distributed around 0 – that’s normal. Histogram is close to normal distribution. Q-Q plot closely follows the straight line, which is good, and Correlogram shows 1 significant value.

* 1. **One step ahead forecast**

The next task is to make a one step ahead prediction using last ten observations. ARIMA model is fitted setting 1 step as a condition. Next 10 last data points taken from the dataset to make a prediction. Prediction calculated and this is the result in a form of plot:

A graph with red lines and numbers

Description automatically generated

*Figure9. One-step-ahead prediction plot*

The blue data point in the graph is the predicted value. The X-axis displays indexes, not the exact Date. It is calculated as 9416.49. The Mean Absolute Error (MAE) is calculated, and it is 3687.50.

1. **Visualisations**

The next step is to create series of visualisations, focusing on the age group of 18-35 year olds. To do this panel dashboard is created using both static and interactive visualisations.

* 1. **. Data preparation for visualisations**

The target audience for this analysis is younger adults aged 18-35. Dataset contains variable Age\_Group, making easier to separate this particular group. For this purpose, original two age groups “Young Adults (<25)” and “Adults (25-34)” are merged into one group “Younger Adults (18-35)”. Also, customers, that are 35 are being moved to this group to. After this is done, new dataframe df\_1835 is created containing just data for this customer segment.

* 1. **Interactive dashboard**

The dashboard chosen to visualize data representing 18-35 age group, and aimed to analyse insights for this age group is Panel. This dashboard is easy to use, read and understand. It works by adding layers one by one, each layer is grouped by the themes:

1. First layer shows the general comparison off all three age groups and displays how significant 18-35 age group is comparing to the other groups.

To display the size of the 18-35 age group in the dataset simple pie chart is used. It shows, that this group of younger adults is the largest, taking up more than a half of all the customers and this group should be the main target of the marketing campaigns of the company. Older adults take up almost 46%, and seniors just 0.6 percent – it’s a very insignificant group.

Alongside the pie chart is displayed interactive bar chart. Starting position is stacked barchart, so the user can see how different age groups are compared across different countries. Below is a dropdown menu, enabling user to further analyse distribution in different countries. Colours of the bars are kept the same within the countries for easier reading, as when the colour changes, user knows, country has changed too.

1. Second layer starts displaying data just about 18-35 age group and shows total revenue generated by this group worldwide. The map is fully interactive and animated, showing how spending of younger adults changed over the years.
2. Third layer displays sales performance in all three categories. It shows that Accessories and Clothing categories were only introduced in 2013 and despite sales being steady since their launch day, are still no match for the main selling category – Bikes. Sales lines also suggest that dataset is suitable to perform machine learning models. The best fitting regression lines can be implemented.
3. Fourth layer shows fully interactive heatmap combined with interactive bar plot. Colours used here are Instagram based, something 18–35-year-olds can relate to. User can use bar plots bars by clicking on them to look at the quantities of items sold.
4. The fifth and last layer shows the sunburst graph. Graph gives a quick idea what items are popular amongst 18–35-year-olds, what sells well and what not so well. To get a more detailed look, graph offers the possibility to look at exact figures using hovering tool. Graph is filled with bright youthful colours, giving great contrast between categories and sub-categories.

Panel dashboard is now complete. Is made about younger adults and for younger adults. Stereotypically this group is more technically savvy, often on the move and like to access information on their handsets. Panel dashboard provides this opportunity acting as an app. Dashboard can be viewed within the notebook or in the new window.

**Conclusion**

Dataset is fully analysed and here are the main findings:

* Collaborative filtering suits well for retail customers;
* Content filtering suits well for retail customers using “Quantity” as pivot table rows, but using “Sub\_Category” as rows, suit better to wholesale customers;
* - Dataset is stationary according to finding of Augmented Dickey-Fuller test and ARIMA model is implemented;
* One-step-ahead forecast using last 10 observations is calculated;
* Interactive Dashboard created capturing the main points of 18–35-year-old customers.

**References and used sources**

Miesle, P., 2023. *https://www.datastax.com.* [Online]   
Available at: https://www.datastax.com/guides/what-is-cosine-similarity  
[Accessed 22 05 2024].

**Word Count**: 2332

**GitHub link**

https://github.com/CCT-Dublin/data-visualization-pt-ca2-ZygimantasJakubauskas