Assignment 3: Predicting Future Outcomes

Nai-Wen, Chung

LSE_DA301 Assignment Report

My Github Link: https://github.com/feinwc/LSE_DA301_assignment

Assignment activity 1: Make predictions with regression

Turtle Games would like to understand customer behaviour and predict sales to create more business success. To help Turtle Games to understand the relationships between loyalty points (dependable variable) and age, remuneration, and spending scores (independent variables), three simple regression models were built. The result shows that spending scores and remuneration have a positive relationship with loyalty points, meaning the higher the spending score or the remuneration a customer has, the higher the loyal points the customer acquires. Compared to remuneration, spending scores seem to have a stronger association with loyalty points (Adj. R-squared: 0.452), explaining that proximately 45% of loyalty points can be explained by spending scores. Approximately 38% of loyalty points can be explained by remuneration (Adj. R-squared: 0.379). Interestingly, age has a negative relationship with loyalty points as the coefficient is showing as minus (-4.012805), indicating that the older a customer is, the fewer the loyalty points the customer acquires.

Assignment activity 2: Make predictions with clustering

To understand the segmentation of the customer in Turtle Games, the clustering model k-means was applied to identify groups with the given data. Firstly, the scatterplot and a pair plot were created to understand the distribution of the dataset. The result shows that most data points are showing as overlapped. However, it is still easy to identify 5 clusters in the plot. Interestingly, education was input as the hue, but the result shows that it seems not the ideal attribute to cluster the dataset. Afterwards, the elbow and Silhouette methods were investigated to improve the accuracy and define the k cluster number. The outcome demonstrates that 5 clusters are the ideal cluster number, which aligned with the distribution of the scatterplot and the pair plot. Lastly, the visualisation was built. Based on the attributes of each cluster, we can identify new market segmentation (listed as below). The marketing team can the information to target future customers and implement marketing campaigns accordingly.

- Cluster Black: low remuneration and high spending scores
- Cluster Yellow: low remuneration and low spending scores
- Cluster Red: middle remuneration and spending scores
- Cluster Blue: high remuneration and high spending scores
- Cluster Green: high remuneration and low spending scores

Assignment activity 3: Analyse customer sentiments with reviews

To design the future campaigns, Turtle Games downloaded the review from their website and applied NLP to analyse customer reviews.
The re-processing was adapted before the sentiment analysis including dropping unnecessary columns, checking missing values, changing to lower cases, replacing punctuations, dropping duplicates, removing alphanumeric characters and stopwords and tokenising. The process helps to ensure accuracy and increase the efficiency of

sentiment analysis afterwards. Based on the display of the word cloud and the frequency distribution, 15 most common words in both review and summary columns are demonstrated. The text such as “game, great, fun, like, play, good” clearly explains that the customers tend to have positive sentiments towards the game. Interestingly, the review of polarity and sentiment shows that most of the text in the review and summary tend to be natural as most of the text has a sentiment score close to 0. Through reviewing the top 20 positive and negative reviews for both columns, the details of the text provide a deep understanding of customer sentiments. For instance, some customers buy a particular board game as a gift. In this case, the marketing can promote the game during Christmas to increase sales. The negative sentiments can help to spot the pinpoints and the team should implement an improvement plan for the product or service. For instance, text such as “boring and disappointing” and “bad quality” explain that the product doesn’t meet customer needs. The team should take action towards the improvement shortly before losing more customers.

Assignment activity 4: Visualise data to gather insights

The summary, min, max, and mean functions were applied to understand the descriptive statistics of the data. Based on the scatter plots, the three regions tend to have a slight negative relationship between product ID and sales, explaining that the greater the ID is, the lower the sales are. This is interesting as customer behaviour seems to have a similar trend across the three regions. Product ID smaller than 5,000 seems to be more popular than product ID greater than 5,000. The histograms and plot boxes demonstrate that the distribution is skewed across the three regions. There are some visible outliers and 50% of the customers in NA spend approximately 0.5K to 3.1K, the customers in EU spend only 0.4 to 2.1K, whilst global customers spend around 1K - 6K. The result shows that more marketing campaigns that can be considered in Europe and other regions are highly potential as they contributed a significant portion of global sales.

Assignment activity 5: Clean, manipulate, and visualise the data

To compare global sales with North America (NA) and Europe (EU), the aggregation function was applied. The result displays that the customer preference in NA and EU are similar to the global customers as the product ID smaller than 5,000 are both more popular. The histograms and box plots demonstrate that the purchasing power of the customer in NA and EU can be encouraged. Also, the finding explains that Eastern markets can be highly potential as they also contributed to global sales. The deeper investigation of customer behaviour and the enhancement/improvement of marketing campaigns are worth to invest to penetrate more markets. The result of the Q-Q plots and the Shapiro test explains the data is normally distributed. However, the skewness is greater than 3, showing that the data is high and positive skewed (right-skewed) and heavy-tiled(leptokurtic). Both figures for kurtosis are quite high (greater than 3), indicating that the data has heavier tails than a normal distribution. The correlation coefficient (of -0.61 in global sales and of -0.56 in NA+EU sales) suggests negative correlations. Finally, two scatter plots were applied to understand the relationship between product ID and sales. It’s interesting to find the global trend is quite similar to the trend in NA+EU regions, which explains that the customer may have similar customer behaviour, preference or attitudes towards Turtle games. The marketing

team can execute mass marketing strategies to acquire more customers in different regions to extend the market.

Assignment activity 6: Making recommendations to the business

The result of the linear regression model of NA and Global sales (model1) shows that approximately 84% of the global sales can be affected by NA sales (Adjusted R-squared: 0.8385), demonstrating NA sales have a strong relationship with global sales. Approximately 72% of global sales can be influenced by EU sales (model2, Adjusted R-squared: 0.7185). Approximately 38% of EU sales are associated with NA sales (model3, Adjusted R-squared: 0.382). The correlation score also indicates the variables have strong correlations with one another ($r \approx 1$). Also, the scatter plots explain that each variable has a positive relationship with one another. The outcome of the multiple linear regression model reveals that approximately 97% of global sales can be explained by NA and EU sales (Adjusted R-squared: 0.9664). NA and EU sales show the significance associated with global sales (***, p-value < 0.001). Through the prediction model, global sales can be predicted by inputting NA and EU sales. The result shows the predicted global sales is ≈ 68.1 when NA and EU sales are equal to 34.02 and 23.80 respectively; ≈ 7.3 when NA and EU sales equal to 3.93 and 1.56 respectively; ≈ 4.9 when NA and EU sales equal to 2.73 and 0.65 respectively; ≈ 4.8 when NA and EU sales equal to 2.26 and 0.91 respectively; ≈ 26.6 when NA and EU sales equal to 22.08 and 0.52 respectively. In summary, NA and EU sales are highly associated with global sales in this model. Although the model has performed well, more variables can still improve the model to create business success in real life.