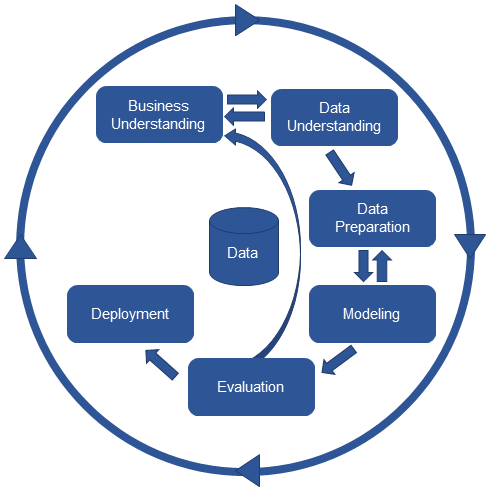
**Introduction**

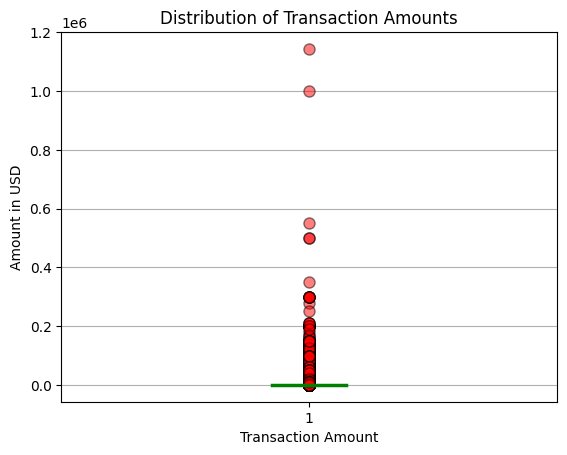
Customer lifetime value or CLV provides an estimate of the revenue or profit that can be expected from an average customer; hence, it is a vital metric in business as it also shows the company’s market value. The projects task was to analyze the data gain insights to advise the company, model the CLV, and create a predictive model that can estimate the CLV of each customer within a specific time window (Caldwell, 2021).

**Methodology**

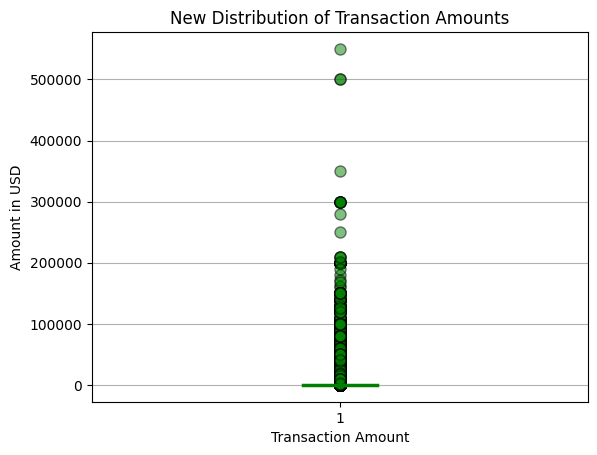


The methodology involved importing necessary libraries and reading the raw customer data from a JSON file. I then checked for missing values on the dataset followed by identifying and removing transactions before the FTD. In general, the methodology involved data cleaning, feature engineering and finding the CLV for every customer.

**Data analysis**



As the above results of the boxplot shows that the majority of transaction amounts are concentrated around the lower end of the distribution of transaction amounts, there were also some extreme values of customers who either withdrew or deposited huge amounts. These extreme values had a significant impact on the mean and standard deviation of the transaction amounts that could have affected the CLV value. So as to have a better analysis of the dataset, I handled the extreme values or outliers by removing them (OpenStaxCollege, 2013).



Even after removing the two extreme values, the data was still concentrated on the lower distribution thus showing the company’s customer base is comprised of mostly of customers who engage in smaller transactions.

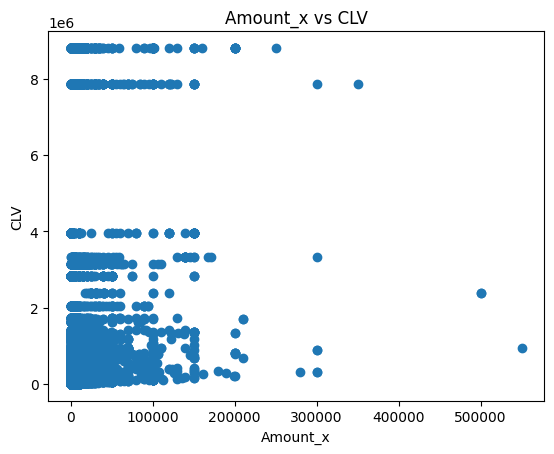
Afterwards, I calculated the CLV that displayed clearly some customers were more valuable to the company than others. Hence, the company could use the CLV information to prioritize customer service efforts towards high-value customers, providing them with premium service to ensure their continued loyalty. Furthermore, the company could use the information to tailor their customer retention and acquisition strategies.

|  |  |
| --- | --- |
| Correlation coefficient | 0.31336063392295666 |
| : p-value: | 0.0 |

On investigating the relationship between the amount of a transaction and the CLV of the customer, I used Pearson’s correlation and obtained the p-value as shown above. The results indicated a positive between the two variables as the p-value showed the relationship was statistically significant and did not happen by chance.

As per the analysis, it appeared to be a moderate positive correlation between the amount spent by customers and their CLV. That highlighted customers who spend more had the highest CLV due to their increased loyalty or higher purchasing frequency. The company should therefore, explore ways to incentivize customers to spend more, such as targeted promotions to increase the customers’ loyalty and CLV.

Though the relationship between the amount of a transaction is significant, it is also not a linear relationship as seen on the scatter plot below, most customers are concentrated on the low CLV henceforth the company could develop ways to increase the CLV of its customers.



I then examined the relationship of the status and type of the transaction.

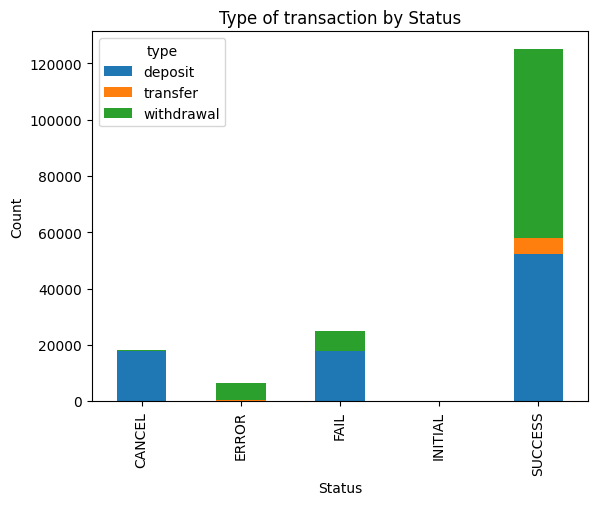
Contingency table

|  |  |  |  |
| --- | --- | --- | --- |
| type | deposit | transfer | withdrawal |
| status |  |  |  |
| CANCEL | 17887 | 0 | 350 |
| ERROR | 104 | 415 | 5955 |
| FAIL | 17786 | 0 | 6971 |
| INITIAL | 6 | 0 | 1 |
| SUCCESS | 52162 | 5883 | 67065 |

Chi-square statistic: 31504.279565713685

P-value: 0.0

The chi-square test results suggest that there were a statistically significant relationship between the status and type variables and the p-value of 0.0 showed the relationship was not by chance. The contingency table above showed most transactions fell under SUCCESS status and deposit transaction type. Moreover, there was a higher proportion of "withdrawal" transactions that result in an "ERROR" status compared to "deposit" and "transfer" transactions (Frost, 2017).



That is also visible on the bar pot above implying that the success rate of a transaction could be influenced by its type. As the bar graph shows, the majority of the transactions are successful, with deposit and withdrawal transactions having a higher success rate compared to transfer transactions.

Based on the above analysis, I would therefore suggest the company to increase focus on successful deposit transactions because as the analysis shows, majority of transactions were successful deposit transactions. The company could consider investing in improving the user experience for this type of transaction to make it even more streamlined and convenient for customers.

Additionally, the company could work on reducing the error rate for withdrawal transactions. That is because as the analysis revealed, there were a significant number of errors in withdrawal transactions. The company should consider to investigating the root causes of these errors and take steps to reduce their frequency. They should include improving the functionality of the withdrawal process and providing more on training the customer service representatives to address withdrawal issues.

Besides, they could consider expanding the product offering as though majority of the transactions are made up of deposits and withdrawals on the company’s business. There was also a substantial amount of transfer transactions. The company should explore expanding its product offerings to include additional types of financial products and services that may be of interest to customers who are using the transfer service (SEC & UNITED STATES - SECURITIES AND EXCHANGE COMMISSION, 2010).

The company should also monitor transaction data progressively as the analysis proof that the transaction data can provide valuable insights into the company's business operations. Therefore, it is recommended that the company continue to monitor transaction data on an ongoing basis to identify trends and patterns that would develop insights on the areas to improve as company. The company could prefer doing that using automated data analysis for accurate and fast insights; they could also decide to use regular reviews by the company's data analysts.

**Model**

The model was first preprocessed using the same steps used during the data analysis of the raw data. That was followed by removing duplicated transactions, I then split the dataset into training and test sets where the features were separated from the CLV variable where hot encoding was performed on the categorical variables.

I then used the Random Forest Regressor to train the model on the training set after saving the model, I loaded it back to the environment where I used it to make predictions on the new customer data that was preserved for evaluating the model. Finally, the results were evaluated using R-squared and Root Mean Squared Error (RMSE) where scatter plot was also generated to visualize the predicted and actual values.

The model chosen, Random Forest Regressor is a supervised learning algorithm that builds a decision tree ensemble. It is an ensemble learning method that creates a robust and accurate model from multiple decision tree models. The algorithms achieves that by creating a set of decision trees from randomly selected subsets of the training data, and then aggregates the results of each tree to make a final prediction.

I chose the algorithm also because it is a powerful algorithm that can handle both classification and regression problems as in regression the algorithm has proved to predict a continuous target variable based on a set of input features. Furthermore the algorithm works by first selecting a random subset of features from the available input features followed by, building multiple decision trees using those selected features and the corresponding target variable. The trees are then constructed by recursively splitting the data into smaller subsets based on the selected features, with the goal of maximizing the homogeneity of each resulting subset in terms of the target variable (E R, 2021).

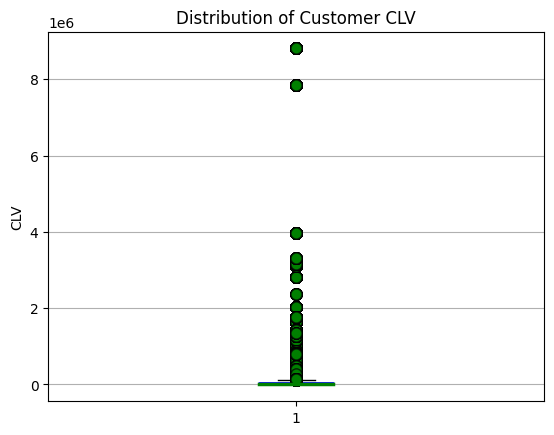
Moreover, when making predictions the Random Forest Regressor algorithm aggregates the predictions of all the trees in the forest. That is accomplished by taking the average of the predicted values for regression problems, or by selecting the most frequent predicted class for classification problems.

Finally, Random Forest Regressor has numerous advantages over other regression algorithms as it is not prone to overfitting than single decision trees because it combines the predictions of multiple trees. It is also able to handle a large number of input features, and can identify the most important features for making predictions (E R, 2021).

**Results**

After transforming the data to timeseries I did a summary statistics that resulted to some insights. Based on the summary statistics the company had 171,938 transactions with a mean transaction amount of 1,040.31 units and a standard deviation of 6,919.27 units. The minimum transaction amount was 0 while the maximum transaction amount was 550,000 units.

The CLV variable seemed to have a wide range of values, with a mean of 138,379 units and a standard deviation of 635,476.6 units as the maximum value being 8,807,684 units indicated that some customers are highly valuable to the company.



As was the case for transactions amount, the CLV amount was also widely spread showing there was small number of customers who were highly valuable to the company leaving most of the customers having low CLV.

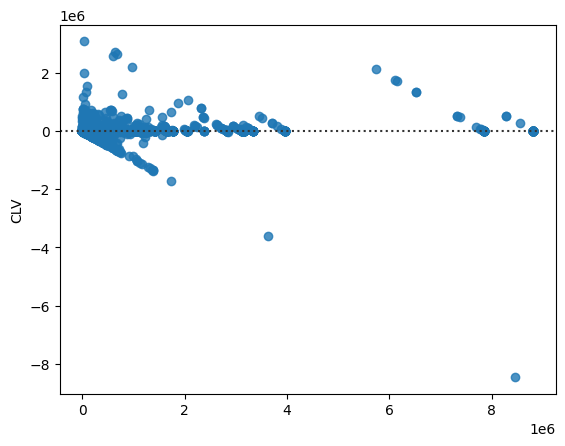
**CLV Model evaluation**

The model had an R-squared value of 0.98 that showing the model explained a very high value of 98% of the spread in the test data. Thus indicating the choice of model was a good fit for the data and had a high predictive probability. However, this could not be the case on the data I used as the data was non-linear hence a high R-squared does not necessarily guarantee the model’s predictive ability as it usually assumes the model was used on a linear dataset (Fernando, 2021).

The model also had a root mean squared error (RMSE) of RMSE: 80408.52. The root mean squared is a measure of the difference between the predicted and actual values in a regression analysis model. Provided the value of the root mean squared on the model it signified that the average difference between the predicted and actual customer lifetime value was around $80,408.52.

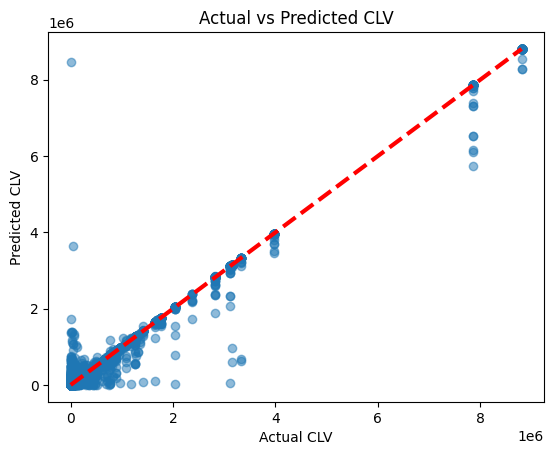
A root mean squared of $80 408.52 is averagely a very large difference highlighting in contrast of the model’s R-squared value that the model may not be a good fit for the data. Nevertheless, because the CLV range was also very high, the root mean squared could be acceptable for the model and be accurate enough for practical use. This however should be backed by the results of residual plots to confirm.

I then used the residual plot to check for homoscedasticity, normality and linearity. The residual plots show the results below. Most residuals are randomly scattered around the horizontal line zero suggesting the model captured most important information. The model also seems to perform better in some ranges of predicted values than others furthermore, the are normally distributed around zero suggesting the model should be appropriate for the data.



**Results**

The experimental results showed that the machine learning model used for CLV prediction has a root mean squared error (RMSE) of 80408.52, that indicated that the predicted CLV values have an average error of $80,408.52 compared to the actual CLV values. The baseline method using the average past transaction number and value had an RMSE of 179954.89, that is more than twice as large as the RMSE of the machine learning model suggests that the machine learning model was significantly better at predicting CLV than the baseline method. Also, the scatter plot below shows a moderately positive correlation between the predicted and actual CLV values because most of the points are clustered around the red dashed line.



The experiment results above also show that the model had a high level of accuracy for predicting the future CLV value for customers, with an R-squared value of 0.98. The comparison of the model again shows that the model outperforms the baseline method, as it is able to make more accurate predictions of the future CLV. The scatter plot above of the predicted and actual values shows a strong positive correlation, with most of the points falling close to the diagonal line. That specifies that the model is able to accurately predict the CLV for most of the customers in the test set.

Lastly, the results suggest that the model is a promising tool for predicting the future CLV of customers, that can help the company make data-driven decisions to optimize customer retention and maximize revenue.