SMU SINGAPORE MANAGEMENT UNIVERSITY

Grab

SMU-X

Project Statement

Grab is a technology company that has grown into SEA's leading car-hailing, ridesharing and food delivery service. Apart from Singapore and Malaysia, Grab has presence in many other countries in Asia such as Vietnam, Indonesia, etc.

Grab would like to have an audit analytics solution as part of their GrabRewards system, more specifically to establish if points are used or expire according to the products' T&Cs and highlight accounts with unusual allocation patterns using machine learning.

Our Approach

1. Implement a second layer of error-checking to identify risks that are not picked up by current preventive measures.
2. Utilize unsupervised machine learning techniques to identify patterns and identify risks within the GrabRewards process.
3. Utilize predictive analytics to form expectations, allowing high risk items to be scoped in during the audit of GrabRewards.

MODELLING

1 Basic Analytics

Identifying conditions that can serve as a check to ensure that the system is functioning as it should. Several predefined conditions were obtained from the data dictionary , while logical conditions had to be determined from the dataset. Various checks were performed on those conditions to ascertain whether the system is functioning as intended. Checks were done for within given datasets and across datasets.

Checks with identified anomalies were highlighted for further investigation for Grab.

Example

Within datasets:

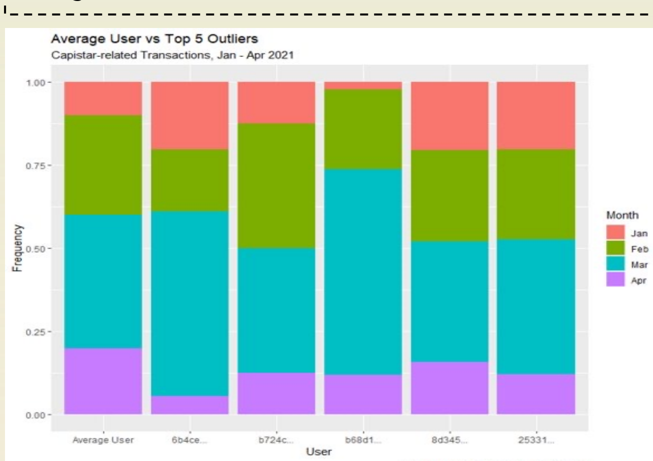
- Check: whether the created time is before the start time, because a record should be created in the system before the start of a reward
- Check whether the number of points that the reward was purchased for is the same as the points displayed in the catalogue

Across datasets:

- Check that the redeemed time of the reward is within the start and end date of the voucher, since the reward can only be redeemed within its validity period

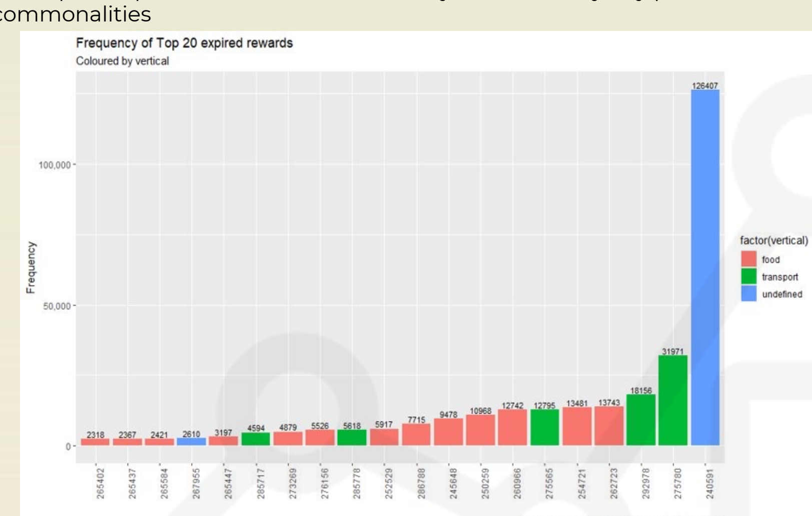
2 Clustering

Principle Component Analysis (PCA) was performed as a dimension reduction technique to reduce the number of variables. K-Modes clustering was the chosen algorithm due the data received being categorical in nature.



Outlier analysis was done, where the top 5 users were identified in the outlier group. They were the biggest spenders, with majority of transactions relating to Capistar vouchers. An average user spends around 5,000 points, but these top 5 users have spent from a range of 33,000 to 96,000.

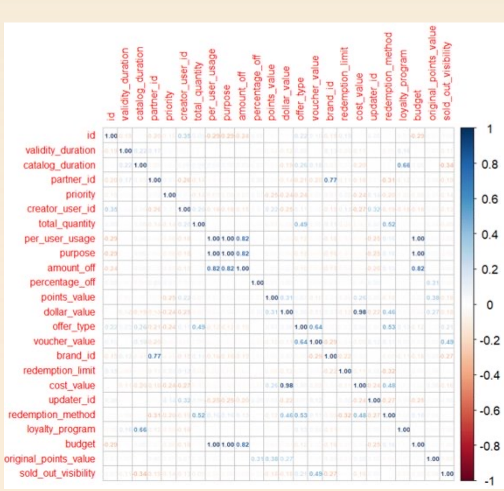
The top 20 expired rewards were also analysed to identify any patterns or commonalities



The top 20 rewards are either technology-related rewards, or driver-related rewards with campaign codes that had similar code structures. Reasons for such expiration of vouchers may suggest a large number of rewards were not redeemed, which may be due to disinterest or ignorance.

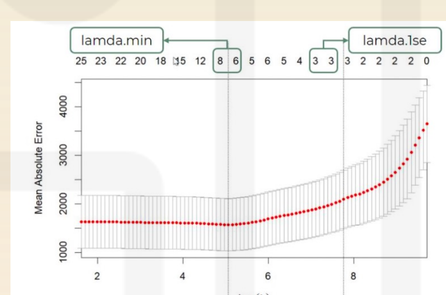
3 Regression

A regression model was created to predict the number of redemptions per GrabRewards such that the predictions can be used as a basis to form audit expectations. Any large deviations from the model can be deemed as high risk items that require greater attention and should be included in the audit scope



After some data cleaning and data manipulation, Exploratory Data Analysis was performed to have a better understanding of the relationships between the IVs and the DVs. Variable selection tools such as LASSO were used to deal with multicollinearity.

Mental models were created and data partitioning was performed. Various regression models were then built to see which one had the best results



Lamda.min and lamda.1se for Elastic Net Regresstion

Various regression models were then built to see which one had the best results

	lm_pred1	en_pred3	XGB_pred	rf_pred
lm_pred1	1.0000000	0.9880271	0.4403538	0.9544994
en_pred3	0.9880271	1.0000000	0.4737956	0.9843729
XGB_pred	0.4403538	0.4737956	1.0000000	0.4649331
rf_pred	0.9544994	0.9843729	0.4649331	1.0000000

Ensemble techniques were used to get the best of all the models. The correlation between each model found that most models were closely related to one another, besides XGBoost.

XGBoost was the most accurate model with the lowest MAE