Rider Loyalty Modelling

Varadraj R Poojary

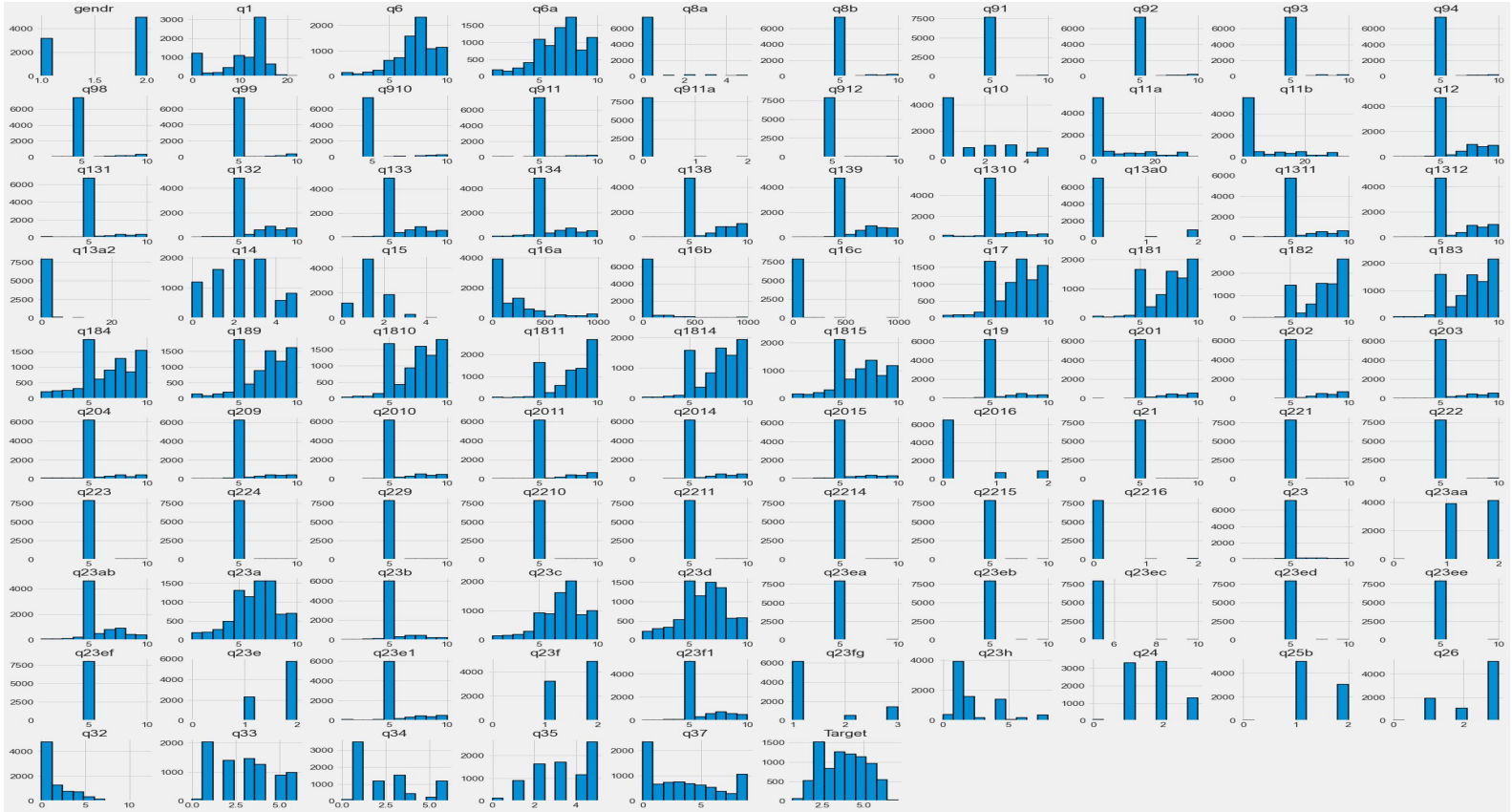
Background & Objective

-The data for this analysis came from a Translink poll of bus, SeaBus, and SkyTrain riders to determine how satisfied they are with the transportation service.

-Respondents are asked about their transport usage (frequency, modes, trip purposes, and time of day), followed by a series of questions about their "satisfaction" with the transit's service quality.

-The goal was to create a loyalty metric for modelling purposes and determine which survey items had an impact on riders' or customers' loyalty.

Feature and target set distributions:



Loyalty Metric

- The number of years were converted to months and added to the number of months to create a loyalty statistic based on how long the consumer has been taking transportation on a regular basis.

- As a weight, added the value for the question: "How likely are you to continue to take transport as frequently as you do today in the near future?" to the value above.

- As a weight, added the value for the question "How likely would you be to suggest Greater Vancouver's transit service to a friend?" to the value above.
- To get a normal distribution, converted the value to log scale.

Evaluation Metrics

- Root Mean Squared Error is the square root of Mean Squared error. It measures the standard deviation of residuals.

- The coefficient of determination or R-squared represents the proportion of the variance in the dependent variable which is explained by the model. It is a scale-free score i.e. irrespective of the values being small or large, the value of R square will be less than one.

- The mean absolute percentage error (MAPE) is a measure of how accurate a forecast system is. It measures this accuracy as a percentage, and can be calculated as the average absolute percent error for each time period minus actual values divided by actual values.

Linear Models

Ridge	1.906 (+/- 0.009)	0.044 (+/- 0.004)	-1.235 (+/- 0.015)	-1.130 (+/- 0.004)	0.131 (+/- 0.033)	0.273 (+/- 0.007)	-33.407 (+/- 0.870)	-30.511 (+/- 0.215)

- Ridge is a simple linear model that fits quickly and produces accurate results.

- As can be observed from the scores, ridge does not do well on our data because it does not appear to follow a linear trend.

- Ridge is performing better than the dummy regressor

- The best reported alpha after hyper parameter optimization is 1 and the best r2 score is 0.1307, which is the same as the default model.

Non-Linear Models

Ridge	2.052 (+/- 0.017)	0.050 (+/- 0.012)	-1.234 (+/- 0.015)	-1.130 (+/- 0.004)	0.131 (+/- 0.033)	0.273 (+/- 0.007)	-33.406 (+/- 0.868)	-30.510 (+/- 0.215)
random forest	147.820 (+/- 1.686)	0.650 (+/- 0.170)	-1.211 (+/- 0.014)	-0.451 (+/- 0.002)	0.163 (+/- 0.025)	0.884 (+/- 0.002)	-32.782 (+/- 0.709)	-12.084 (+/- 0.065)
XGBoost	8.061 (+/- 0.136)	0.039 (+/- 0.004)	-1.237 (+/- 0.009)	-0.755 (+/- 0.015)	0.128 (+/- 0.014)	0.675 (+/- 0.012)	-32.826 (+/- 0.585)	-19.069 (+/- 0.400)
LightGBM	1.960 (+/- 1.408)	0.059 (+/- 0.012)	-1.198 (+/- 0.013)	-0.903 (+/- 0.003)	0.182 (+/- 0.015)	0.536 (+/- 0.005)	-32.310 (+/- 0.739)	-24.186 (+/- 0.142)
CatBoost	11.581 (+/- 0.171)	0.078 (+/- 0.014)	-1.194 (+/- 0.009)	-0.870 (+/- 0.004)	0.187 (+/- 0.019)	0.569 (+/- 0.003)	-32.251 (+/- 0.692)	-23.238 (+/- 0.145)

- Random forest averages trees, but boosting-based models grow by learning from past trees' mistakes.

- Tree-based and boosting-based models outperform the dummy and the ridge(linear) model

- The CatBoost model has the best test r2 of 0.189.

Overfitting/Underfitting:

- Because the difference between the train and test scores is so large, the Random forests model overfits.
- All other models seem to be performing fine on the data.

Fit time:

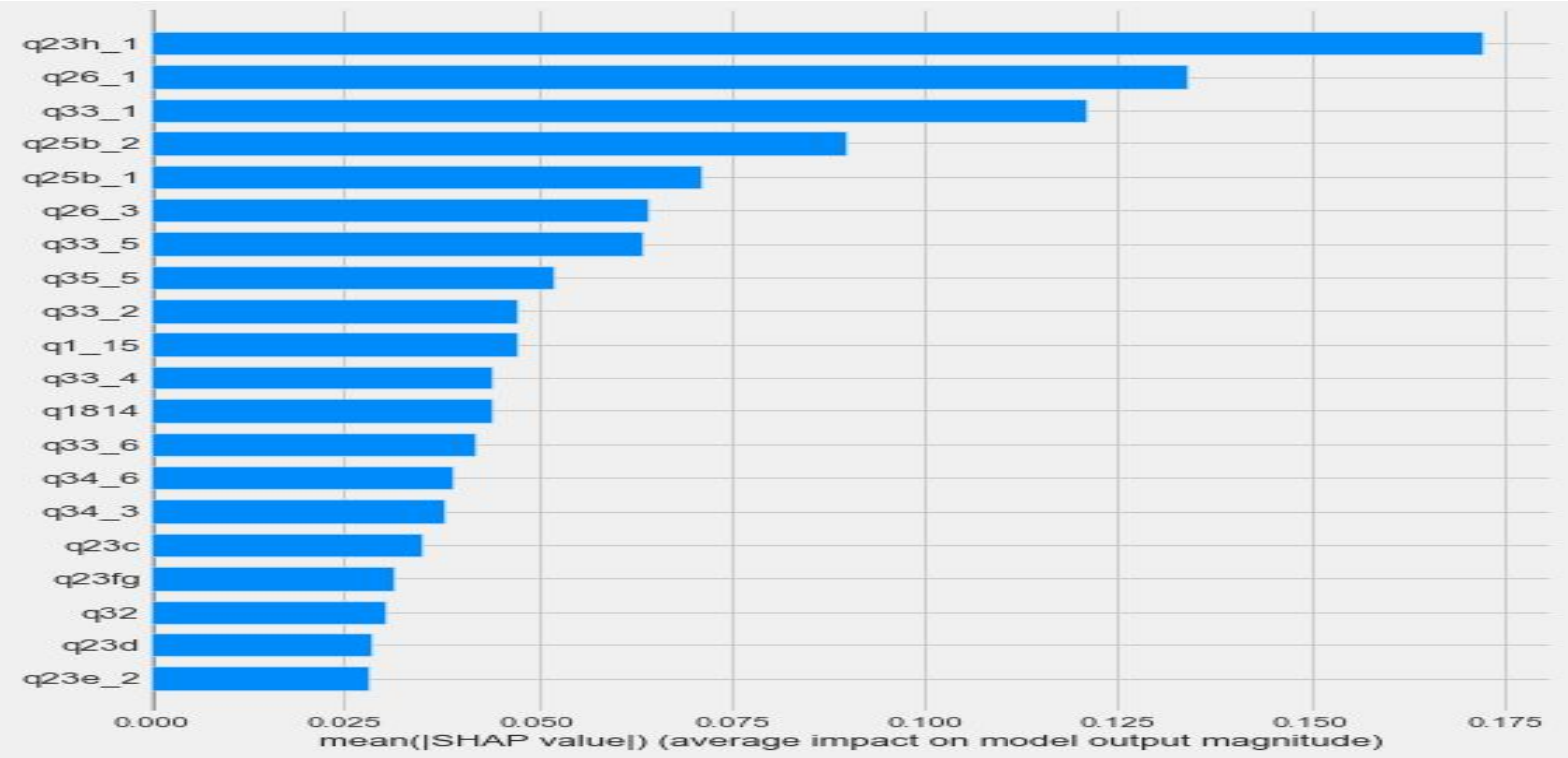
- The Random forest takes the longest to fit because the maximum depth isn't specified and the model overfits the data.
- LGBM has the lowest fit time compared to all other models.

Score time:

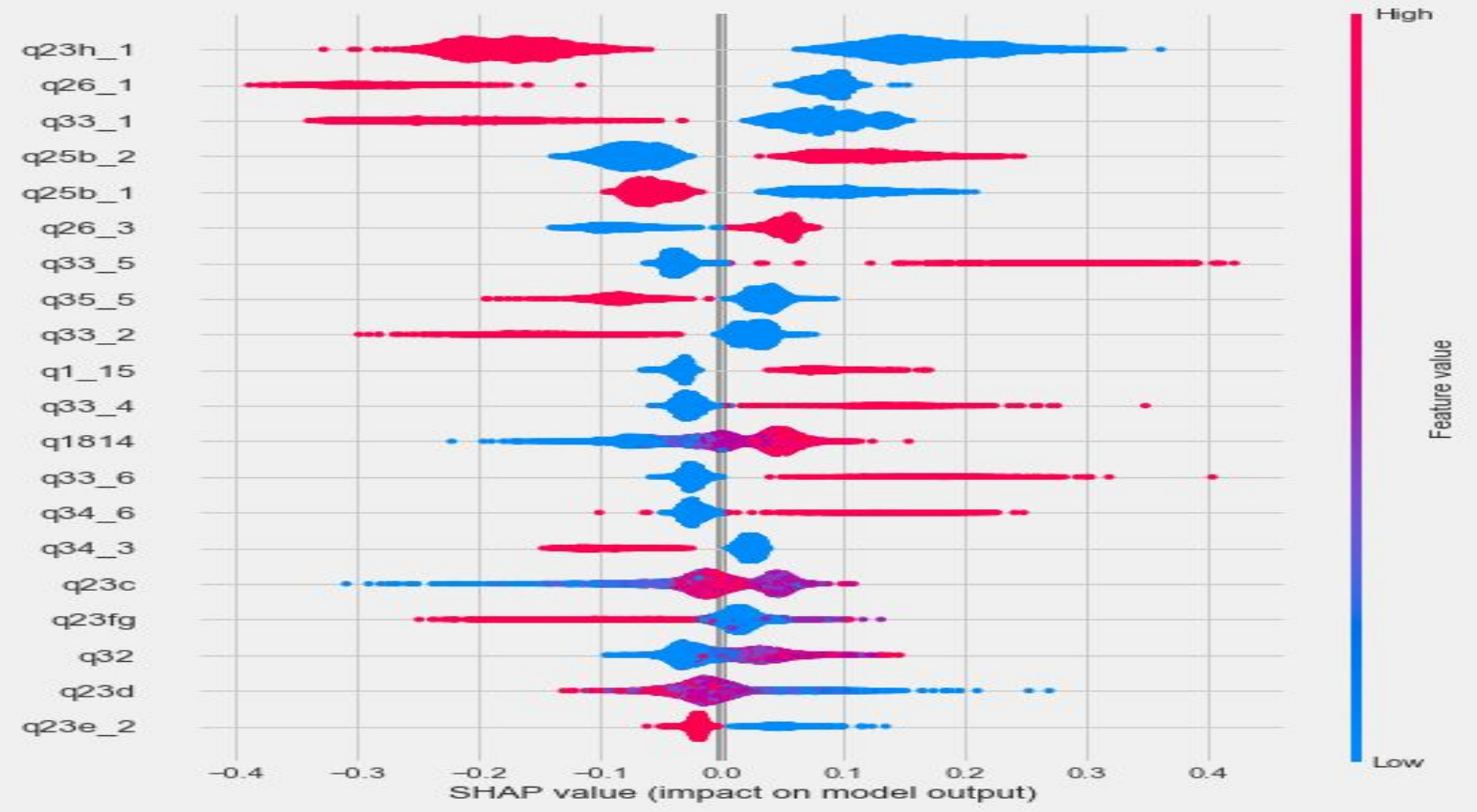
- Random forests has the highest score time.
- LGBM has the lowest score time compared to all other models.

Features of Importance

Plot 1:



Plot 2:


q23h_1
q26_1
q33_1
q25b_2
q25b_1
q26_3
q33_5
q35_5
q33_2
q1_15
q33_4
q1814
q33_6
q34_6
q34_3
q23c
q23fg
q32
q23d
q23e_2
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
SHAP value (impact on model output)
High
Feature value
Low

Plot1:

The average SHAP value for each feature is shown in plot 1, or the first plot. The feature 'q23h' with value 1, i.e. the use of cash as a payment method, has the biggest impact on the predictions, according to the plot, while the feature 'q26' with value 1, i.e. the feature representing new consumers, has the second highest impact.

Plot2:

Plot 2 depicts the importance of each feature as well as the direction in which that feature influences the prediction.

Top 5 interpretations from both the plots:

1) Customers are more loyal since they use less cash as a mode of payment.
2) The loyalty score of new customers is lower.
3) Customers who are older are more loyal.
4) People without access to personal transportation or who do not possess a car are more loyal.
5) Compared to six months ago, customers who are riding transit about the same are more loyal.

A prediction for a single test set value is shown above.

- As can be observed, the higher value/presence of cash payment drives the prediction in negative direction.

- The lack of access to a car or other personal transportation modes drives the prediction in negative direction.

- Being in the 45-54 age group drives the forecast in a favourable direction, implying that older customers are more loyal.

Results

Model	Type	Dataset	Test R2 score	Test mape	Remarks
Dummy Regressor	Non-linear	Train	0.001	37.55	Baseline
Ridge	Linear	Train	0.131	33.4	Liner/simple model, less fit time ,low r2
Random Forest Regressor	Non-linear	Train	0.165	32.75	Low mape score, Overfitting, high fit time
XGBoost Regressor	Non-linear	Train	0.129	32.81	High mape score, low r2
LGBM Regressor	Non-linear	Train	0.182	32.31	Less fit time, low r2
CatBoost Regressor	Non-linear	Train	0.189	32.23	Best individual model,significant fit time

Using the survey data from Translink , started with the problem statement of predicting a loyalty score for the data points.

Performed the given steps.

1) Data cleaning: According to my understanding, the data was changed or enhanced. Features with no questions and those that were unrelated were removed. Unrealistic target data was removed. The data contained an excessive number of Nan values, which were addressed with.

2) EDA: Obtaining the connection between features and determining which features should be engineered and preprocessed/transformed in which manner.

3) Loyalty Metric : Created a new loyalty metric based on 4 other features from the dataset.

4) Preprocessing: Preprocessed our features according to their kind and what was learned from the EDA.

5) Different Models: Tried out different regression models on the train data i.e. The Baseline, Linear and Non Linear Models.Compared them based on scores, fit and score time and Underfit or Overfit. Got the Best model as CatBoost based on the scores.

6) Feature Selection: Using lasso and select from Model, tried feature selection and looked for improvements in cross val scores. There was no use in considering feature selection because there was no improvement.

7) Hyper parameter optimization: On the non-linear models ,performed Hyper parameter optimization and checked for improvements in scores.

8) Interpretation and feature importances: Implemented shap on our best model i.e. CatBoost to gain insights as to which features have major impact on our prediction and which direction they drive the predictions to. From the given step we are able to infer the feature 'q23h' with value 1, i.e. the use of cash as a payment method, has the biggest impact on the predictions, while the feature 'q26' with value 1, i.e. the feature representing new consumers, has the second highest impact.

9) Test data Evaluation: We used test data to evaluate our trained best model and presented the results as well as our knowledge of the predictions.

Executive Summary

- The features 'q23h' with value 1, i.e. the usage of cash as a payment method, and 'q26' with value 1, i.e. the feature indicating new customers, are important in predicting a customer's loyalty.

- Translink can use these characteristics to anticipate a customer's loyalty in advance.

- Also, use these features to provide a better user experience and focus on a certain user group.

- Alternatively, improve the experience for the users who were left out.

Future Scope / Enhancements

- May have attempted stacking or averaging the regressors to improve the scores.

- The data may have been changed to incorporate polynomial features in order to create a linear model.

- A better feature selection method can also be utilised.

Thank you