Feature analysis and total delivery time modeling report

The problem

Predicting total delivery duration in seconds is a regression problem. The project prompt explicitly states that "underprediction of delivery time is of particular concern as past experiments suggest that underestimating delivery time is roughly twice as costly as overestimating it". One of the ways to solve this problem is to provide confidence intervals for the predictions along with the most likely predictions.

Looking at the data and generating new features

New features has been added to the dataset based on common sense. For example, delivery time should be affected by month, day of the week, and time of day, so these features are added. Also, the sum of the estimated order place duration and store to consumer driving duration is added as a new feature. The time in seconds has been computed and added as the target variable for regression.

The data examination process reveals that different types of variables are present in the dataset. The data is imbalanced and contains a lot of missing values. The distributions of variables differ considerably, and the outliers are present (e.g., negative number of dashers). No outliers have been removed in this work, and a robust approach has been used instead. Quick analysis has shown that the features are not suitable for standard methods of dimensionality reduction. The model choice has to account for all the problems described above.

The model

The model of choice is XGBoost with decision trees. In this particular project, XGBoost models with different hyperparameters have been compared to each other, and no other type of models have been used. However, similar process should be followed with other models suitable for this type of data (e.g., neural networks). The advantage of XGBoost is that it is interpretable and provides feature importances that have been used for feature selection in this project. The performance of the current model has been found to be ~25% better compared to the prediction with the median value (no model).

Two additional XGBoost models ("quantile models") have been used for quantile regression to predict confidence intervals for the values predicted with the main model. They produce the confidence intervals that cover the observed values in ~78% of cases.

Finding the most important features

The main model has been used for recursive feature elimination with cross-validation to filter out the most important features for predicting the total delivery time. It has been demonstrated that recursively eliminating 8 features from the dataset does not affect the performance of the model considerably, but the performance is better when all features are used (*Fig.1*). These 8 least important features are (less important to more important): 'num_distinct_items', 'store_primary_category', 'total_busy_dashers', 'estimated_order_place_duration', 'max_item_price', 'order_protocol', 'total_items', 'min_item_price'. The 3 most important features found are (more important to less important): 'created_at_month', 'created_at_hour', and 'market_id'. The full lists of eliminated features at each step may be found in *Attachment 1*.

Interpretation of the results and recommendations

Total delivery time is strongly associated with time and location. However, the existing data only contains information for two months and only morning and evening delivery times. It is recommended to collect more data for these unknown gaps to improve prediction performance. The data that may be associated with time and location, e.g., traffic conditions for each location at the specific time, should be collected.

As total delivery time is strongly associated with the number of orders and the number of on-shift dashers, it is also recommended to collect the data on on-shift dashers, e.g., their 1-5 rates based on reviews. The depersonalized ID of each dasher (encoded names) may also increase prediction performance.

The dataset should be reviewed to correct extreme outliers (e.g., negative numbers of dashers), and more effort should be put into maintaining a reliable dataset.

Only XGBoost models are provided in this report, and it is recommended to research other types of models as well.

Putting the model into production

The model may be evaluated and compared with the existing model using shadow testing (Champion / Challenger approach). If the new model's predictions are significantly better than those of the old model, the old model is substituted with the new one without delay in time. The time period required for collecting the right dataset to establish statistically significant defferences may be estimated beforehand.

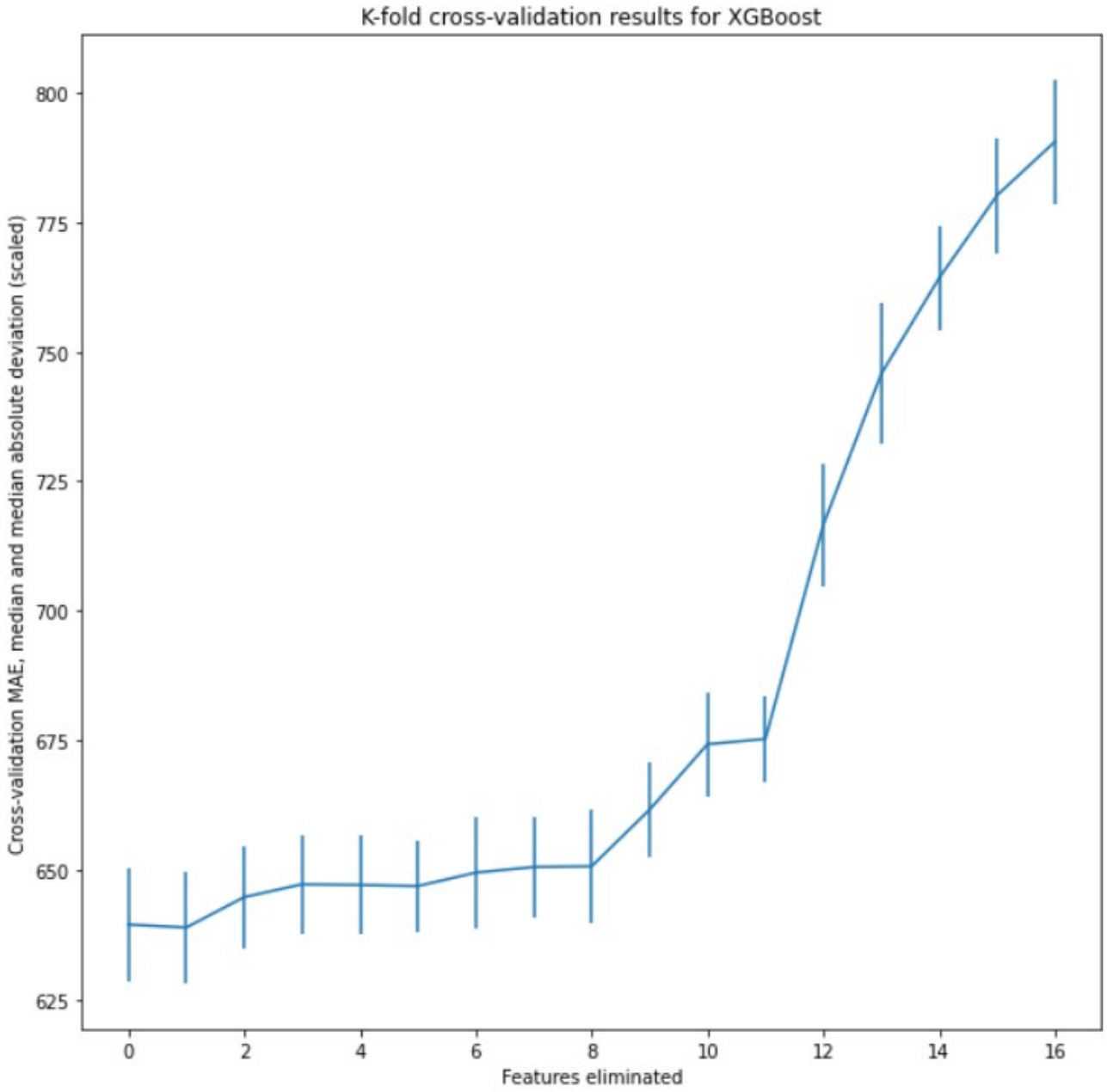

Figure 1: Recursive feature elimination with cross-validation. Eliminating up to 8 features does not affect model performance significantly, but the performance is best with all features included. The data for the plot with complete list of features are provided in Attachment 1.