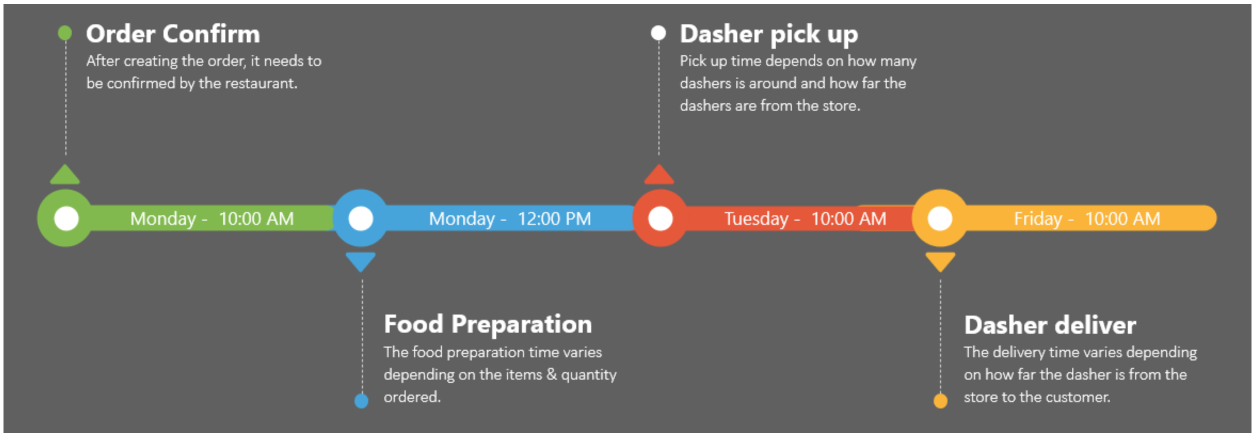
Food Delivery Time Prediction

Project Overview

The goal of this project is to predict the food delivery time which has a big impact on consumer experience. The feed delivery time can be broken down into 4 phases: Order confirm -> Food Preparation -> Dasher pick up -> Dasher deliver.



Order Confirm

After creating the order, it needs to be confirmed by the restaurant.

Dasher pick up

Pick up time depends on how many dashers is around and how far the dashers are from the store.

Food Preparation

The food preparation time varies depending on the items & quantity ordered.

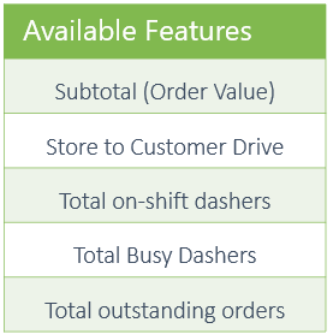
Dasher deliver

The delivery time varies depending on how far the dasher is from the store to the customer.

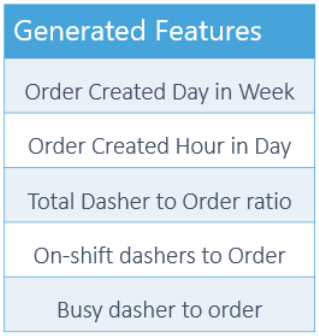
Ideally, each phase should have it's own model given the fact that the drivers are very different. For example, Food preparation time heavily depends on the items & quantity being ordered. And Dasher delivery time depends on the distance from Store to customer. Due to the data limitation, here we are developing a top-line model which directly output the predicted based on the features.

Key Features and Potential Features

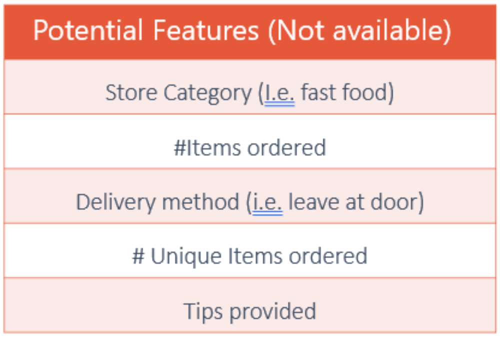
Dasher to Order ratio is generated to capture how fast the food will be picked up. Order created Day and hours are created to capture the seasonality of delivery time. The following table summarized all features that have been explored during the model development and potential features that could significantly improve the model accuracy:



Available Features
Subtotal (Order Value)
Store to Customer Drive
Total on-shift dashers
Total Busy Dashers
Total outstanding orders



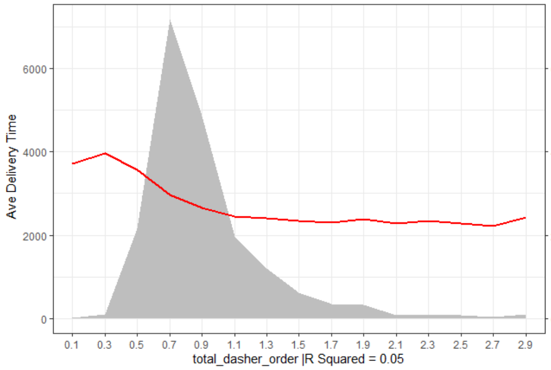
Generated Features
Order Created Day in Week
Order Created Hour in Day
Total Dasher to Order ratio
On-shift dashers to Order
Busy dasher to order



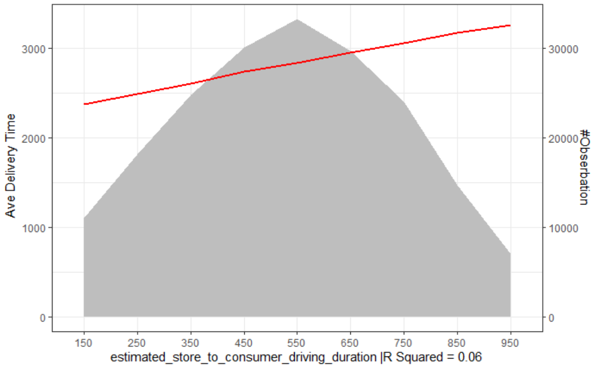
Potential Features (Not available)
Store Category (I.e. fast food)
#Items ordered
Delivery method (i.e. leave at door)
Unique Items ordered
Tips provided

For each available feature, we visualized the empirical relationship between it's value and the delivery time and estimated the univariate importance measure by R^Squred. Here, we only listed the important features and the explanation:

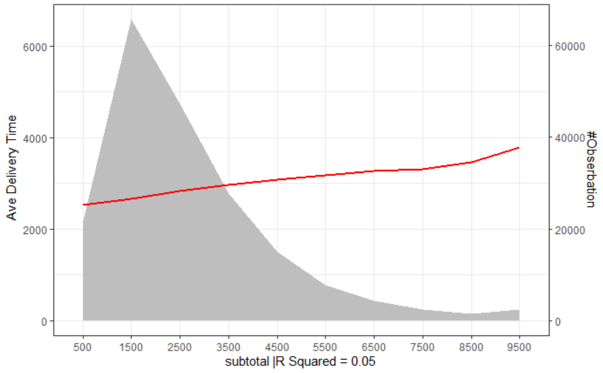
Total Dasher to Order ratio



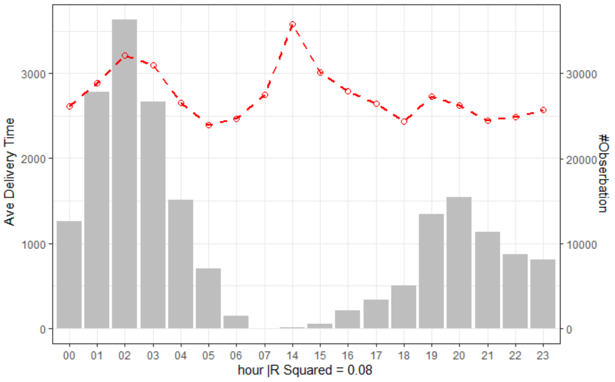
Estimated store to customer duration



Subtotal



Hour in a Day



Total Dasher to Order ratio shows up a negative relationship with delivery time, which is in line with our expectation that the more dasher available for the order, the faster it would be picked up. Subtotal shows up a positive relationship with delivery time, which is caused by the fact the higher value food might require longer time to prepare.

Model Performance

3 models have been tested here: Linear Regression, Random Forest and GBM. In order to test the model performance, a 30/70 split has been performed to sample data such that model performance can be tested in the out of sample dataset.

The matrix we are using to compare the model performance is R Square and RMSE (Root Mean Squared Error). The following table shows that RF and GBM have similar model performance while Linear model is worse than both. We decided to choose Random Forest Model as the champion model because better out-of-sample performance. The matrix is shown below:

	Key Parameter	In sample Rsq	Out of Sample Rsq	In sample RMSE	Out of Sample RMSE
Linear Model	Spline Transformation for key features + Stepwise feature selection	**0.273**	**0.267**	**867.4**	**867.3**
Random Forest	mtry=4; mtree=70 (My personal laptop is not powerful enough to cross validate tune the pamaters with mtree>70)	**0.308**	**0.312**	**847.6**	**841**
GBM	ntree=10000, depth=1, learning rate=0.1	**0.316**	**0.3**	**841.1**	**847.9**

Other Consideration

In order to demonstrate the new model is better the previous one, there are two aspects need to be considered:

1. The prediction accuracy measure by MSE and R Square.
2. The percentage of predictions that are VERY FAR from the actual delivery time.

The first point is very intuitive and is widely used for almost all models. In generalize Minimizing MSE is to decrease the “average” estimation error, however, severely underestimate/overestimate for a small portion could also have negative impact to the business as orders that are very early / late are also much worse than those that are only slightly early / late.

Due to time limitation, we focus on the first point in this project. But the 2nd point should also be carefully considered in practice.