Introduction

The purpose for this project is to predict the estimated time taken for delivery, in order to improve the consumer experience. The feature of the data come from different dimensions including store feature, order features, market features, and predictions from other models. The dataset contains 16 variables and 197428 observations in training dataset. Various machine learning methods are used to make the predictions, including data cleaning, feature engineering, model building and model evaluation. In this project, RMSE was used as criteria to evaluate model’s performance, which is a way to observe the difference between the prediction and actual value. Therefore, smaller RMSE means better model.

Methodology

Feature engineering

Some new features are introduced:

(1) Order variety:

The high order variety could be adding complexity on food preparation, which means longer time taken for delivery. The smaller the ratio is the bigger the variety the order is.

Chart, scatter chart

Description automatically generated

Graph 1: actual time vs order variety

The relationship between actual delivery time and order variety can be shown by Graph1. People can detect clear tendency that when the ratio is smaller the actual time is larger.

(2) average price per item

Normally, fast foods are cheaper, and the expectation of waiting time for consumers are shorter.

Chart, scatter chart

Description automatically generated

Graph 2: actual time vs average price per item

From the graph2, we can detect the actual time has positive relationship with average price per item.

(3) hour and busy hour rank

The hour variable can be derived from the variable “created\_at”, which is the time showing when the order was ordered by consumer.

Table

Description automatically generated Table

Description automatically generated

Table 1: how many orders are placed in specific hour Table 2: the median delivery time on specific hour

After comparing the table 1 with table 2, we can detect normally the more orders are placed, the longer delivery time it will be taken. The introduction for variable busy hour rank is necessary, for example in hour 2, 36976 orders are placed, which is the most among other hours, so we define hour 2 as rank 1 .

(4) Get Dummies for variable “store\_primary\_category”, “market\_id”, “order\_protocol”

Since the data type for variable “store\_primary\_category” is object, we convert it into dummies variables, which are numeric variable that represents categorical data, taking 0 or 1 to indicate the absence or presence of specific category. And also, different categories may have different time ranges on food preparation. The same reasons can be applied the other two variables.

Building Model

For the model, I used weighted average ensemble method. The six models that we chose to ensemble performed relatively better than other models, which means they have smaller RMSE.

Conclusion and recommendations

The RMSE in the test dataset is about 968.64, which represents the differences between values predicted by the model and the actual value. Other features can also be added into the training dataset to improve the accuracy. One is weather feature like rainfall and wind speed. Normally, the delivery time for sunny day will be less than other conditions like raining. Secondly, we can add the latitude and latitude for consumer and restaurant, so we can estimate the delivery time. Thirdly, we can add numbers of waiting order for the restaurant to estimate time on food preparation.