Porter Delivery Time Prediction

Adi, Daniel, and Patrick

What is Porter

"Operator of an online logistics marketplace intended to offer a standardized service for the utilization of trucks. The company's platform permits businesses to book mini trucks/tempo from a fleet of vehicles as per the date and location, enabling businesses to have access to reliable and economical mini truck booking and letting them track their rides in real time." (Pitchbook)

PORTER

Objective & Value Proposition

Objective

- Develop model to help tell users predicted times
- Learn how to improve delivery times

Value Proposition

- We are consultants for Porter and will be creating a useful model that can help them improve their food delivery service

Dataset

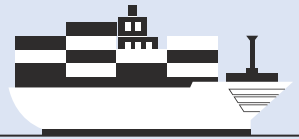
The dataset is from Porter containing restaurant delivery information from 3 weeks

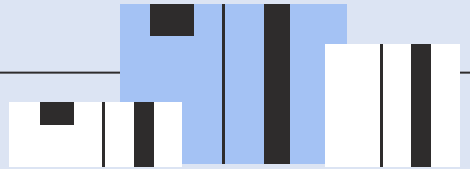
Data Summary:

- 14 columns about relevant delivery information for when each order is placed
- 175,777 unique entries

Column	Datatype
market_id	float64
created_at	object
actual_delivery_time	object
store_primary_category	int64
order_protocol	float64
total_items	int64
subtotal	int64
num_distinct_items	int64
min_item_price	int64
max_item_price	int64
total_onshift_dashers	float64
total_busy_dashers	float64
total_outstanding_orders	float64
estimated_store_to_consumer_driving_duration	float64

EDA

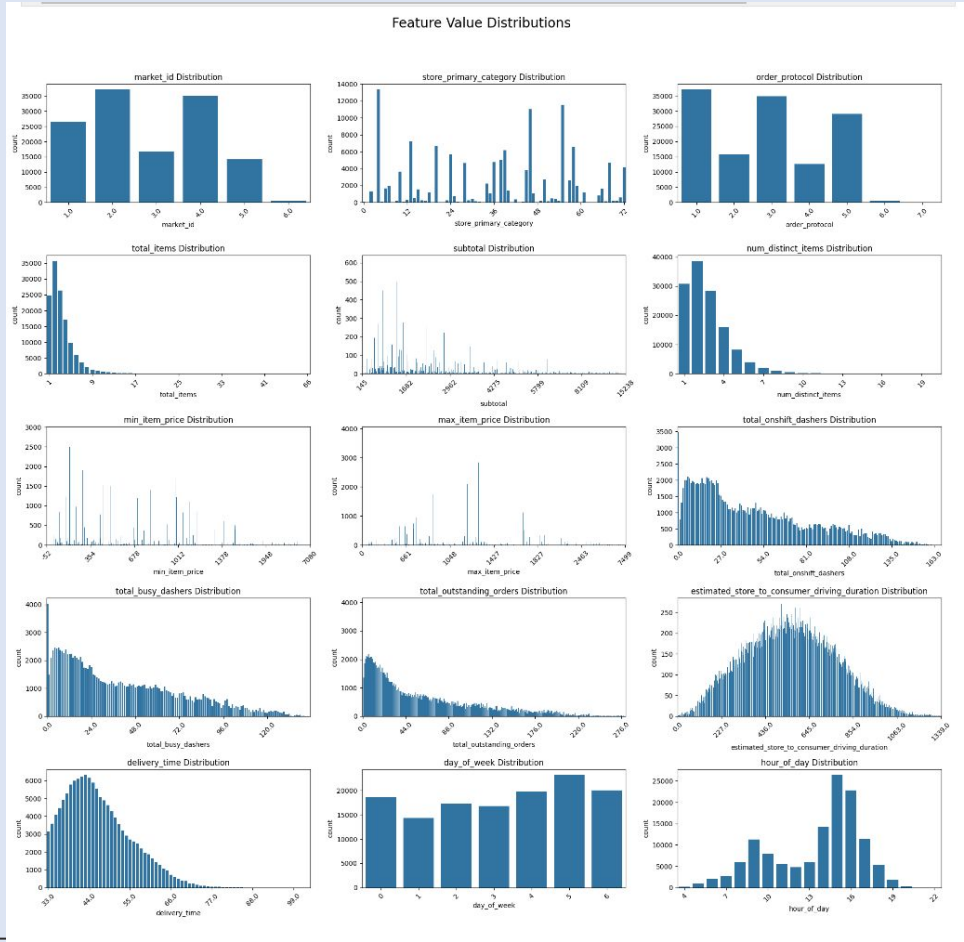




EDA Major Learning 1

Feature Distributions:

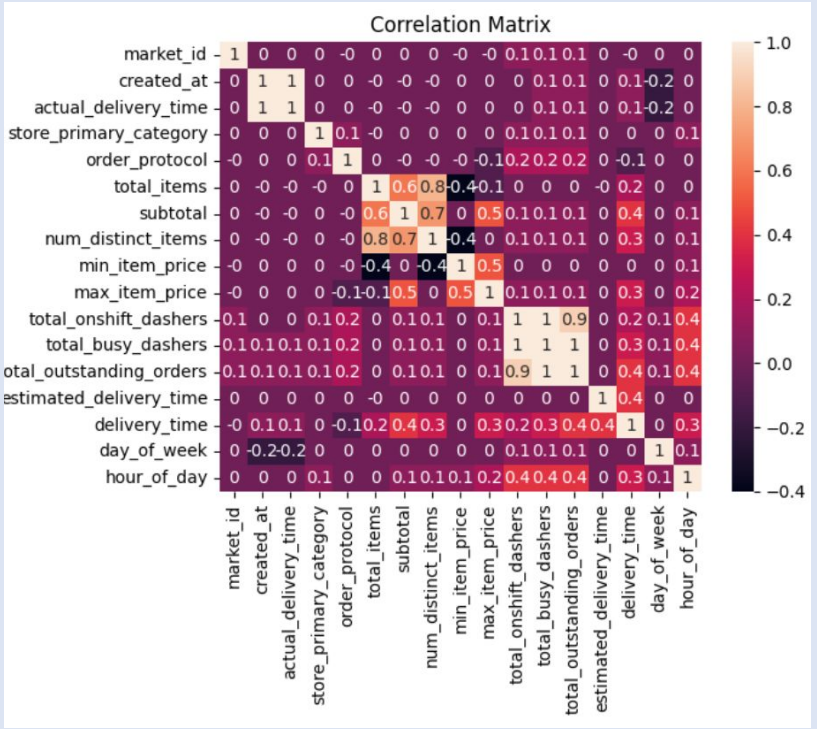
- Market IDs and Order Protocol were generally uniformly distributed
  - Few categories were undersampled
- Continuous features had a right skew
  - Shorter travel times, smaller and cheaper orders
- Delivery time had a normal distribution but removing outlier data smoothed the distribution
  - Negative values in dashers, improper order times, etc.



EDA Major Learning 2

Exploratory Data Analysis:

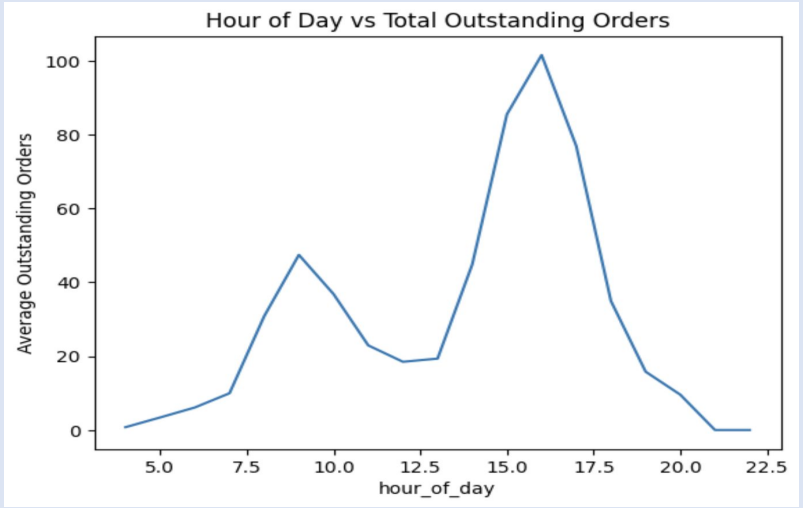
- Uniform delivery times among Market IDs and Order Protocols
  - Avg: ~42 min
- Strong correlation between on-shift drivers, busy dashers, outstanding orders
- Remaining features were uncorrelated

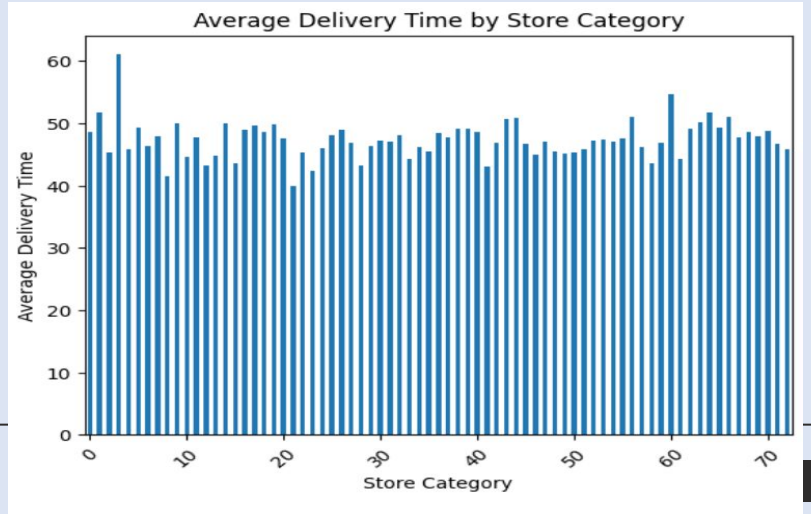


EDA Major Learning 3

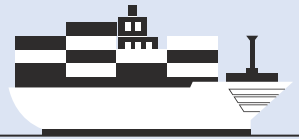
Advanced Feature Analysis

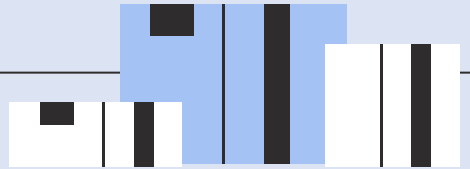
- Higher total orders means longer delivery times
  - Also equated to more onshift dashers being busy.
- No significant correlation between store category and delivery time
  - Uniform delivery time across categories





Modeling Results



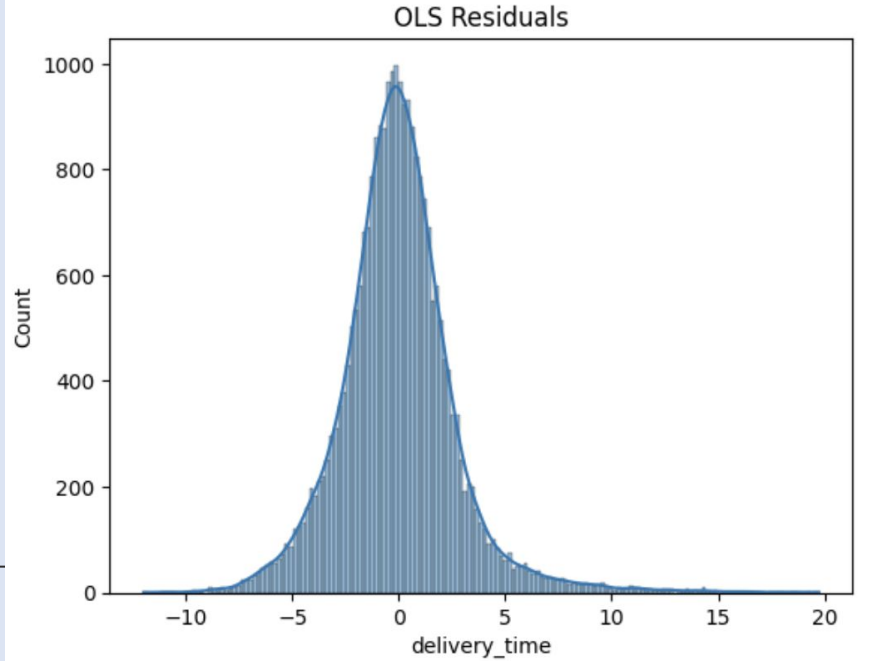


Linear Regression

	coef	std err	t	P>|t|	[0.025	0.975]
total_items	0.0989	0.005	19.586	0.000	0.089	0.109
subtotal	0.0016	7.07e-06	220.542	0.000	0.002	0.002
total_onshift_dashers	-0.2940	0.001	-253.803	0.000	-0.296	-0.292
total_busy_dashers	-0.2208	0.002	-130.603	0.000	-0.224	-0.218
total_outstanding_orders	0.3622	0.001	502.315	0.000	0.361	0.364
estimated_delivery_time	0.0194	4.59e-05	423.854	0.000	0.019	0.020
hour_of_day_a	14.4708	0.067	214.470	0.000	14.339	14.603
hour_of_day_b	18.4064	0.068	270.082	0.000	18.273	18.540
market_id_1.0	8.6507	0.037	232.010	0.000	8.578	8.724
market_id_2.0	3.9749	0.038	104.474	0.000	3.900	4.049
market_id_3.0	4.5000	0.040	111.553	0.000	4.421	4.579
market_id_4.0	4.7058	0.038	124.219	0.000	4.632	4.780
market_id_5.0	5.0013	0.041	122.568	0.000	4.921	5.081
market_id_6.0	6.0445	0.143	42.147	0.000	5.763	6.326
day_of_week_0	6.1611	0.031	199.538	0.000	6.101	6.222
day_of_week_1	3.6567	0.033	110.964	0.000	3.592	3.721
day_of_week_2	3.5465	0.031	112.894	0.000	3.485	3.608
day_of_week_3	3.9618	0.032	125.315	0.000	3.900	4.024
day_of_week_4	4.7913	0.031	156.270	0.000	4.731	4.851
day_of_week_5	4.9400	0.030	167.437	0.000	4.882	4.998
day_of_week_6	5.8198	0.030	192.591	0.000	5.761	5.879
order_protocol_1.0	6.1846	0.104	59.543	0.000	5.981	6.388
order_protocol_2.0	5.4671	0.106	51.747	0.000	5.260	5.674
order_protocol_3.0	4.7288	0.104	45.371	0.000	4.524	4.933
order_protocol_4.0	4.2013	0.106	39.468	0.000	3.993	4.410
order_protocol_5.0	3.2836	0.104	31.436	0.000	3.079	3.488
order_protocol_6.0	4.5792	0.171	26.709	0.000	4.243	4.915
order_protocol_7.0	4.4326	0.727	6.101	0.000	3.009	5.857

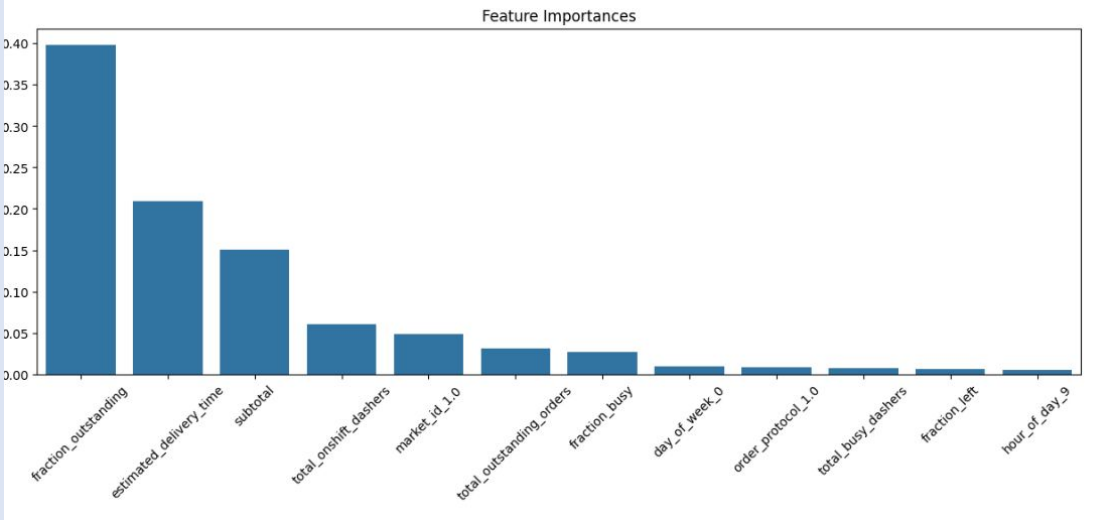
- OLS Linear Regression
  - R^2: 0.9
  - High Kurtosis, Jarque-Bera Test Failed -> non-linear relationships

```
Linear Regression R²: 0.904693365067602
Linear Regression MAE: 1.931482482008211
Linear Regression RMSE: 2.7046361984115763
Linear Regression Kurtosis: 4.329950208674525
Jarque-Bera test statistic: 23599.136972558987, p-value: 0.0
```

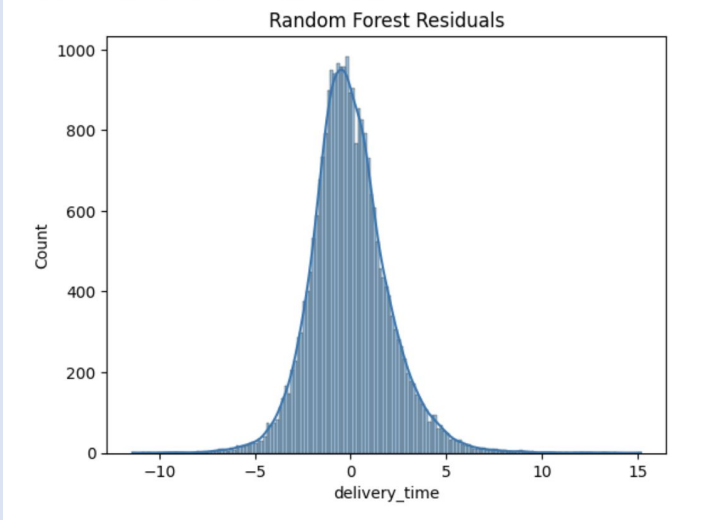


Random Forest

- Random Forest
  - R^2 0.94
  - The residuals were more normally distributed

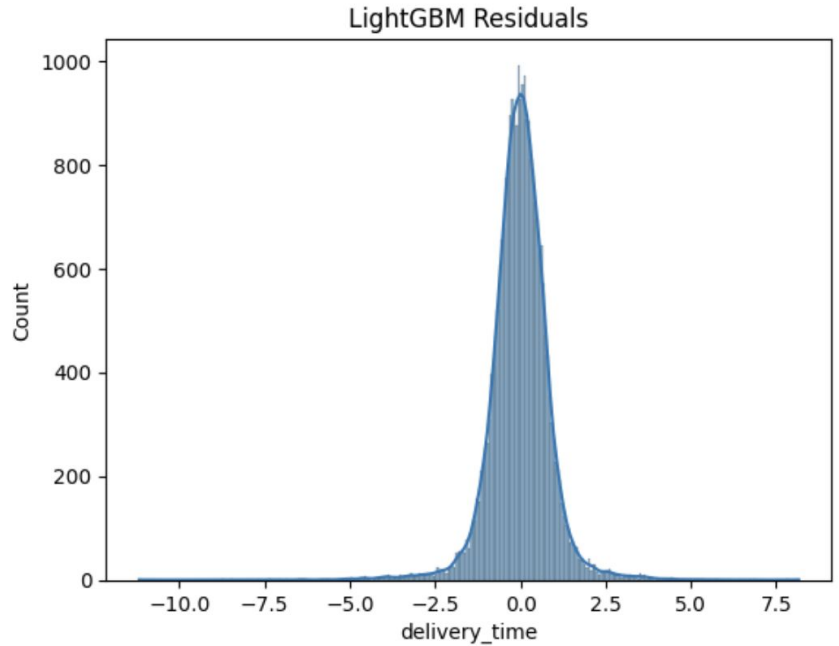


Random Forest R^2: 0.9464741898751157
Random Forest MAE: 1.5297629401577832
Random Forest RMSE: 2.02688408634546
Random Forest Kurtosis: 2.1706315197386337

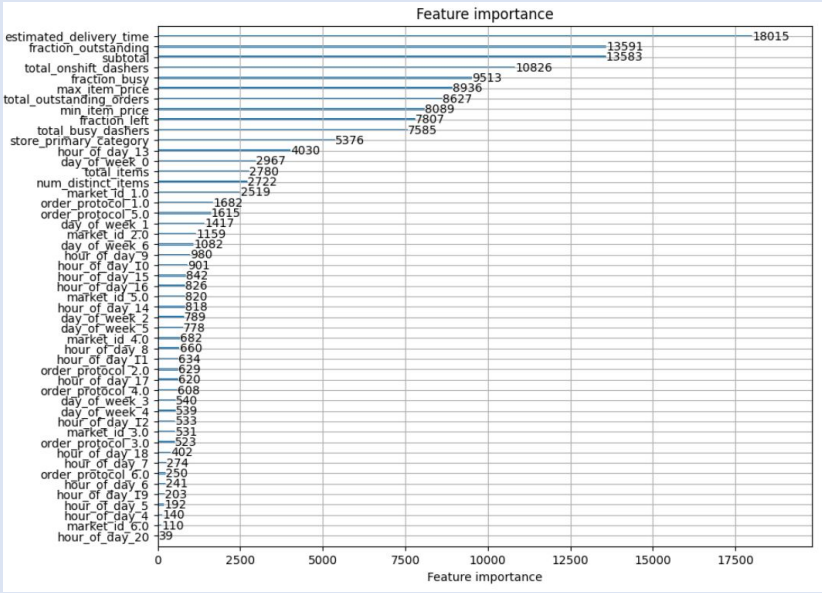


Gradient Boosting

```
LightGBM R²: 0.9900722356127357
LightGBM MAE: 0.6054203938720836
LightGBM RMSE: 0.8729171361348758
LightGBM Kurtosis: 8.721013918347515
```

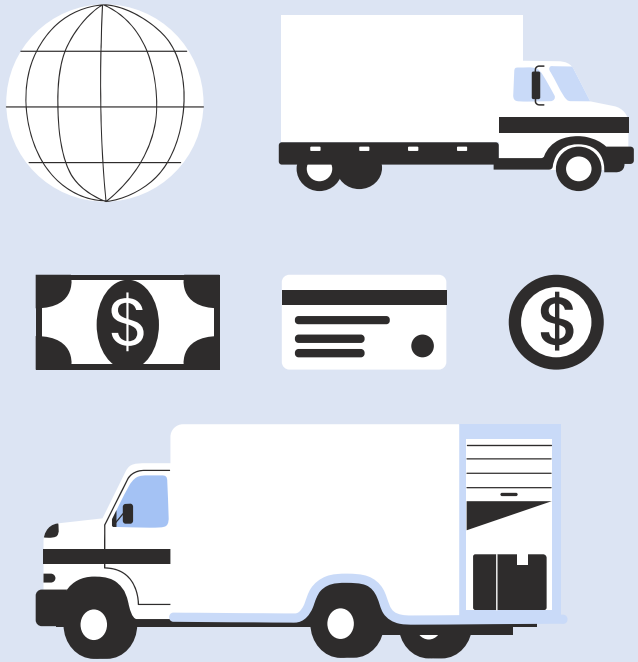


- (LightGBM)
  - R^2: 0.99
  - Impactful features: estimated delivery time, fraction_outstanding, subtotal
  - Most useful model



Implications and Insights

- Feature Importance: Across models, certain features consistently show high importance, such as estimated_delivery_time, outstanding_orders, busy/active drivers, subtotal, and interaction features between these.
- Model Suitability: Our LightGBM model fit our data very well. This is unsurprising, given 100k observations, tabular data, and >>300 features. We achieved an R^2 of .99 which is much higher than ~.8 that one would get with a naive linear regression (before feature engineering).



Challenges and Limitations

Challenges - Parts of our data were potentially misreported. We tried our best to clean the data, but we are unsure if there are still flaws after cleaning.

Limitations - Our final model was unable to accurately predict certain outliers. Even though the LightGBM model performed the best, it still predicted a similar amount of deliveries >10 minutes off.

Potential Future Work - Analyze data over longer term, like 1 year. Consider other factors like environmental features.