Predicting DoorDash Delivery Durations

ADSP 31010 Linear and Non-Linear Models

Team 7: Paul Chang, Irene Liu, Edwin Sanchez-Medina, Yan Xiong

Business Problem

Delivery times estimation is crucial for maintaining customer satisfaction and optimizing operational efficiency in the competitive $353.3 billion online food delivery space

31%

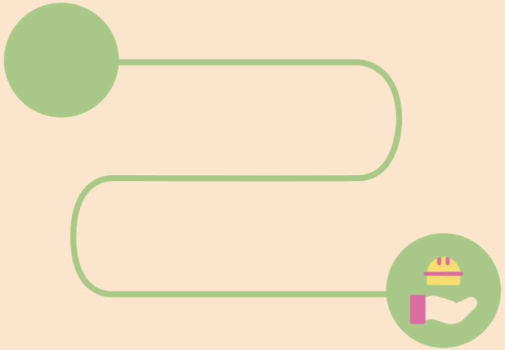
Service Quality
Customer Satisfaction

DoorDash Dataset: Kaggle

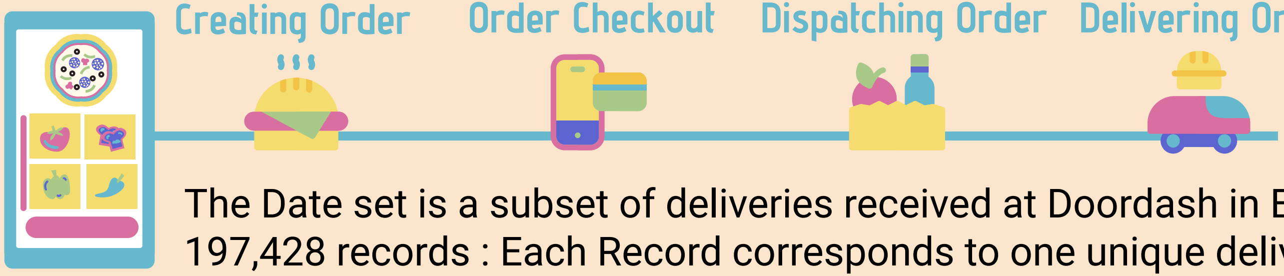
Price — Comparing Prices

Time — Check Estimated Delivery Time

Add-On — Think What Else To Add



Creating Order | Order Checkout | Dispatching Order | Delivering Order



The Date set is a subset of deliveries received at Doordash in Early 2015
197,428 records : Each Record corresponds to one unique delivery

*DoorDash Founded in 2013

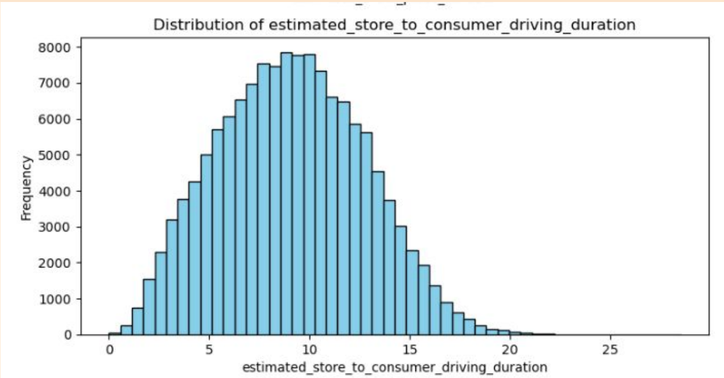
Data Features

- Market_id: float
- Created_at: object
- Actual_delivery_time: object
- Store_id: int
- Store_primary_category: object
- Order_protocol: float
- Total_items: int
- Subtotal: int
- Num_distinct_items: int
- Max_item_price: int
- Min_item_price: int
- Total_onshift_dashers: float
- Total_busy_dashers: float
- Total_outstanding_orders: float
- Estimated_order_place_duration: int
- Estimated_store_to_consumer_driving_duration: float

Feature Creation

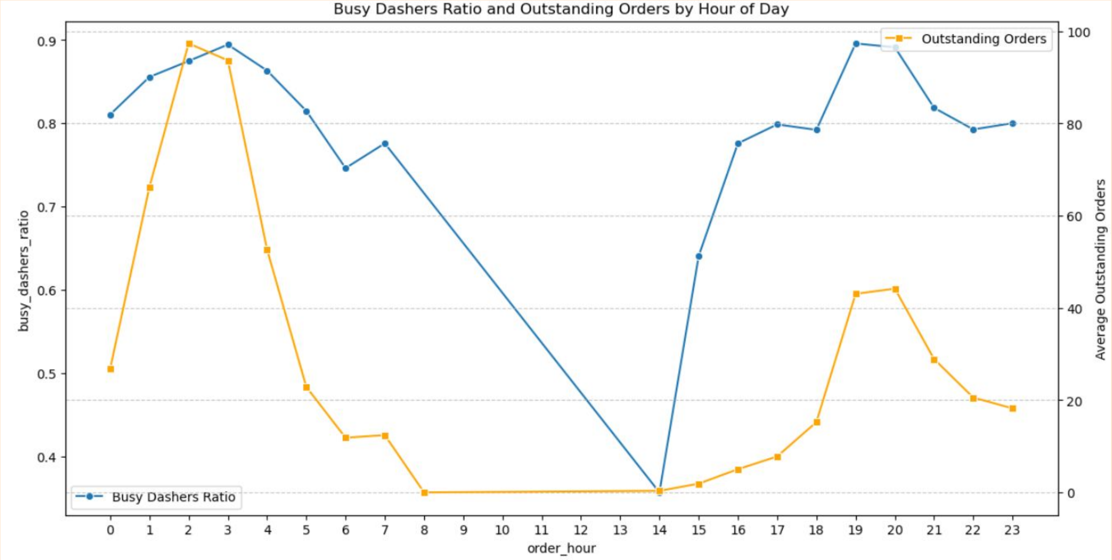
- Busy Dasher Ratio = Total Busy Dashers / Total Onshift Dashers
- Estimated Non-Prep Duration = Estimated Order Place Duration + Estimated Store-to-Consumer Driving Duration
- Filtered data where:
    - Total Busy Dashers ≤ Total Onshift Dashers
    - Minimum Item Price ≤ Maximum Item Price

Exploratory Data Analysis



~ 3 items per order
~ $9 per order

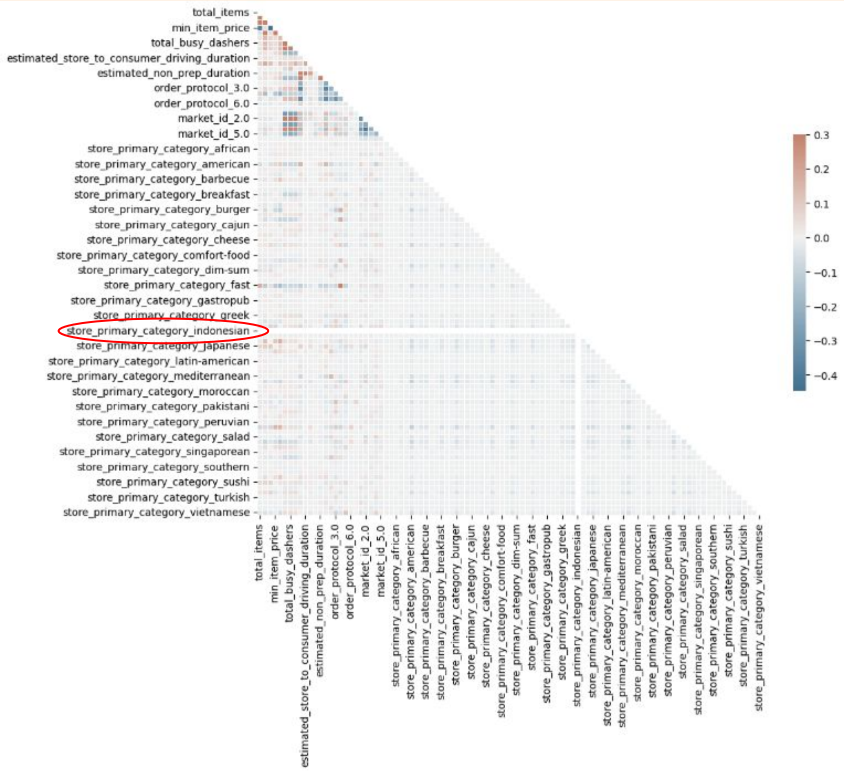
- Estimated **driving duration** from the store to the consumer is approximately **9 minutes**



- Afternoon and Evening (2 PM to 10 PM): The busy dashers ratio spikes in the late afternoon which aligns with common meal times and thus a higher demand for food delivery
- Morning Hours After 2 AM The Busy Dashers stars to decrease

Collinearity

- Applied corr( ) to generate a Correlation Matrix
- Features dropped:
  - Indonesian Store Category; had a lot of 0's as values and hence, no effect



Removing Redundant Pairs

- Top Absolute Correlations
    - Iterative process that identifies the top redundant features in the dataset
- Step 1:
    - Features dropped:
        - Total Onshift Dashers
        - Total Busy Dashers
        - Estimated Non-Prep Duration

Top Absolute Correlations

total_onshift_dashers	total_busy_dashers	0.941741
	total_outstanding_orders	0.934639
total_busy_dashers	total_outstanding_orders	0.931295
estimated_store_to_consumer_driving_duration	estimated_non_prep_duration	0.923086
estimated_order_place_duration	order_protocol_1.0	0.897645
total_items	num_distinct_items	0.758146
subtotal	num_distinct_items	0.682890
total_items	subtotal	0.557175
min_item_price	max_item_price	0.541241
subtotal	max_item_price	0.507947
order_protocol_4.0	store_primary_category_fast	0.489946
num_distinct_items	min_item_price	0.446733
market_id_2.0	market_id_4.0	0.402421
total_items	min_item_price	0.389277
order_protocol_1.0	order_protocol_3.0	0.373581
estimated_order_place_duration	order_protocol_3.0	0.364170
	estimated_non_prep_duration	0.363297
order_protocol_1.0	order_protocol_5.0	0.342345
market_id_1.0	market_id_2.0	0.334580
estimated_order_place_duration	order_protocol_5.0	0.333291

More Redundancy

- Step 2:
  - Features dropped:
    - Order Protocols
    - Market ID’s
  - Additional features dropped:
    - Created At
    - Store ID’s
    - Store Primary Categories
    - Actual Delivery Time

```
Top Absolute Correlations
estimated_order_place_duration  order_protocol_1.0     0.894941
                                order_protocol_1.0     0.894941
total_items                     num_distinct_items     0.746675
subtotal                        num_distinct_items     0.686802
total_items                     subtotal               0.552757
min_item_price                  max_item_price         0.535628
subtotal                        max_item_price         0.508465
num_distinct_items              min_item_price         0.444880
market_id_2.0                   market_id_4.0          0.400374
total_items                     min_item_price         0.381123
estimated_order_place_duration  order_protocol_3.0     0.365637
                                order_protocol_3.0     0.365637
                                order_protocol_5.0     0.330577
                                order_protocol_5.0     0.330577
market_id_1.0                   market_id_2.0          0.309373
total_outstanding_orders        market_id_2.0          0.297235
market_id_1.0                   market_id_4.0          0.296147
total_outstanding_orders        market_id_4.0          0.281039
                                market_id_3.0          0.274844
                                market_id_1.0          0.266156
```

More Redundancy

- Step 3:
  - Features created:
    - Percentage of Distinct Items
    - Average Price per Item
  - Features dropped:
    - Number of Distinct Items
    - Subtotal

Top Absolute Correlations

total_items	num_distinct_items	0.746675
subtotal	num_distinct_items	0.686802
total_items	subtotal	0.552757
min_item_price	max_item_price	0.535628
subtotal	max_item_price	0.508465
num_distinct_items	min_item_price	0.444880
total_items	min_item_price	0.381123
total_outstanding_orders	busy_dashers_ratio	0.216200
	estimated_order_place_duration	0.180989
estimated_store_to_consumer_driving_duration	actual_total_delivery_duration	0.179119
subtotal	actual_total_delivery_duration	0.171507
max_item_price	store_primary_category_italian	0.170541
total_items	store_primary_category_fast	0.164880
max_item_price	store_primary_category_fast	0.164256
	store_primary_category_pizza	0.160654
total_outstanding_orders	actual_total_delivery_duration	0.156857
min_item_price	store_primary_category_pizza	0.153995
actual_total_delivery_duration	busy_dashers_ratio	0.152679
estimated_order_place_duration	store_primary_category_american	0.150561
subtotal	total_outstanding_orders	0.139373

More Redundancy

- Step 4:
  - Features created:
    - Price Range of Items
  - Features dropped:
    - Maximum Item Price
    - Minimum Item Price
- **Total number of features in the dataset went down from 177 to 82**

Top Absolute Correlations

min_item_price	avg_price_per_item	0.858810
max_item_price	avg_price_per_item	0.770937
min_item_price	max_item_price	0.535628
total_items	percent_distinct_item_of_total	0.439205
	min_item_price	0.381123
	avg_price_per_item	0.301566
store_primary_category_pizza	avg_price_per_item	0.230385
percent_distinct_item_of_total	avg_price_per_item	0.224617
total_outstanding_orders	busy_dashers_ratio	0.216200
	estimated_order_place_duration	0.180989
estimated_store_to_consumer_driving_duration	actual_total_delivery_duration	0.179119
store_primary_category_fast	avg_price_per_item	0.175067
max_item_price	percent_distinct_item_of_total	0.173922
min_item_price	percent_distinct_item_of_total	0.172378
max_item_price	store_primary_category_italian	0.170541
total_items	store_primary_category_fast	0.164880
max_item_price	store_primary_category_fast	0.164256
	store_primary_category_pizza	0.160654
total_outstanding_orders	actual_total_delivery_duration	0.156857
store_primary_category_italian	avg_price_per_item	0.156279

Top Absolute Correlations

total_items	percent_distinct_item_of_total	0.439205
	price_range_of_items	0.327488
	avg_price_per_item	0.301566
store_primary_category_pizza	avg_price_per_item	0.230385
percent_distinct_item_of_total	avg_price_per_item	0.224617
total_outstanding_orders	busy_dashers_ratio	0.216200
	estimated_order_place_duration	0.180989
estimated_store_to_consumer_driving_duration	actual_total_delivery_duration	0.179119
store_primary_category_fast	avg_price_per_item	0.175067
total_items	store_primary_category_fast	0.164880
total_outstanding_orders	actual_total_delivery_duration	0.156857
store_primary_category_italian	avg_price_per_item	0.156279
actual_total_delivery_duration	busy_dashers_ratio	0.152679
store_primary_category_fast	percent_distinct_item_of_total	0.150821
estimated_order_place_duration	store_primary_category_american	0.150561
store_primary_category_burger	avg_price_per_item	0.105885
estimated_order_place_duration	store_primary_category_fast	0.105719
actual_total_delivery_duration	price_range_of_items	0.105184
total_outstanding_orders	price_range_of_items	0.104926
store_primary_category_american	store_primary_category_pizza	0.104890

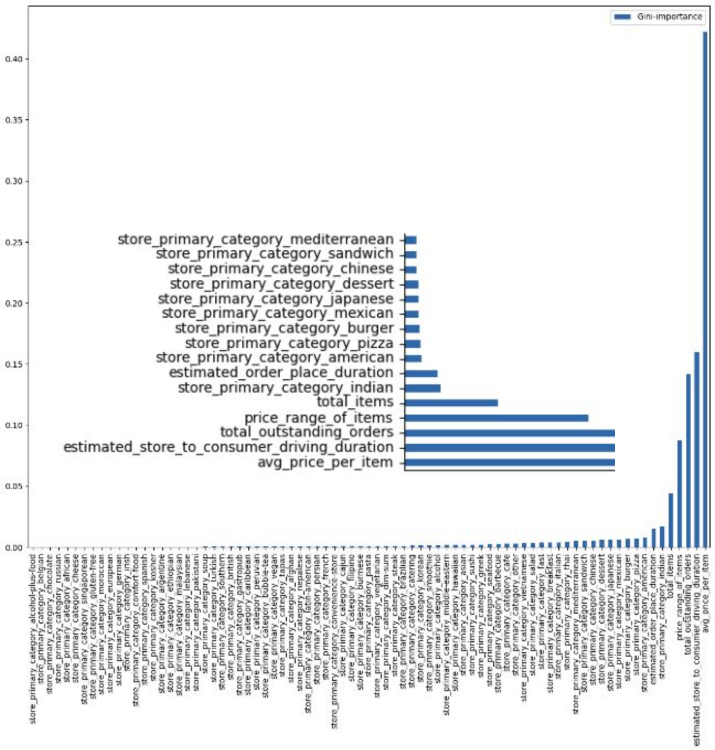
Removing Multicollinearity

- Calculated Variance Inflation Factor Scores to quantify the severity of multicollinearity
- Features dropped:
    - Percentage of Total Distinct Items
    - Busy Dashers Ratio
- **Total number of features in the dataset went down from 82 to 80**

	feature	VIF
0	store_primary_category_alcohol-plus-food	1.000595
1	store_primary_category_chocolate	1.000690
2	store_primary_category_belgian	1.001125
3	store_primary_category_russian	1.003318
4	store_primary_category_african	1.004552
...	...	...
76	estimated_store_to_consumer_driving_duration	7.187204
77	store_primary_category_american	7.824114
78	estimated_order_place_duration	13.495358
79	busy_dashers_ratio	25.402699
80	percent_distinct_item_of_total	30.195754

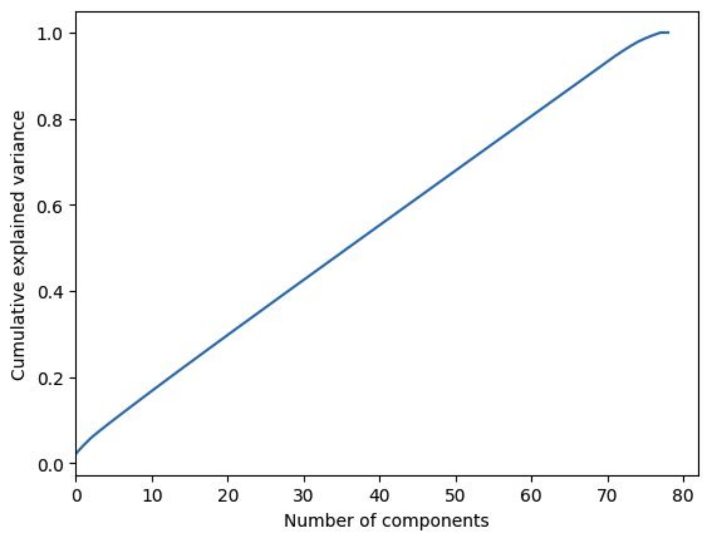
Dimension Reduction

- Want to drop features that do not have a significant effect on the model
- Random Forest:
  - Applied train_test_split (test size = 20%)
  - GINI Index (a.k.a Mean Difference in Impurity [MID])
    - High Index = uniform/impure distributions of target values
    - Low Index = degenerate/pure distribution of target values
  - Sorted features by GINI-Importance



Dimension Reduction & Feature Set Size

- PCA is also effective at eliminating multicollinearity
    - Applied Standard Scalar to the Trained Data (train size = 80%)
- Interpretability:
    - e.g.) ~80% of the data can be explained by using 60 features
- Feature set sizes used for modeling:
    - Full dataset (All 79 features)
    - Top 40 features
    - Top 20 features
    - Top 10 features



Feature Transformation

- Scaling data is important for:
  - Achieving optimal algorithm performances/convergences
    - (e.g. Gradient Descent-based)
  - Ensuring equal importance in features
    - (e.g. Distance-based like KNN, SVM, K-Means)
- Feature scalers used for modeling:
  - Standard Scaler
  - Min/Max Scaler (a.k.a. Normalization)
  - No Scaler

Feature Selection

Feature Transformation	Metric	Full Dataset
Standard Scaler	Train Error	0.9854313731193542 in MLP
Standard Scaler	Test Error	0.7544323205947876 in MLP
Standard Scaler	RMSE	1088.3775634765625 in MLP
Min/Max Scaler	Train Error	0.0043390956707298756 in Ridge
Min/Max Scaler	Test Error	0.0032862715888768435 in Ridge
Min/Max Scaler	RMSE	1091.650146484375 in Ridge
No Scaler	Train Error	1435.70556640625 in MLP
No Scaler	Test Error	1082.99365234375 in MLP

Machine Learning Regressors

- Ridge
- Decision Tree
- Random Forest
- XGBoost
- LGBM
- MLP

- RMSE was still high among all implemented models
- Room for improvement in feature engineering
- Created another feature = Preparation Time = Actual Total Delivery Duration - Estimated Store to Consumer Duration - Estimated Order Place Duration

Feature Selection

Feature Transformation	Metric	Full Dataset	40 Features
Standard Scaler	Train Error	0.9848687052726746 in MLP	1.0090850591659546 in MLP
	Test Error	0.7509970664978027 in MLP	0.7678903341293335 in MLP
	RMSE	1083.4217529296875 in MLP	1087.0216064453125 in MLP
Min/Max Scaler	Train Error	0.004339095670729875 6 in Ridge	0.003137047768214732 3 in DecisionTree
	Test Error	0.0032862715888768435 in Ridge	0.003285221875592390 8 in DecisionTree
	RMSE	1091.650146484375 in Ridge	1091.022180736562 in DecisionTree

Feature Transformation	Metric	20 Features	10 Features
Standard Scaler	Train Error	1.0111454725265503 in MLP	1.0149521827697754 in MLP
	Test Error	0.7691430449485779 in MLP	0.7682734131813049 in MLP
	RMSE	1088.794921875 in MLP	1087.5638427734375 in MLP
Min/Max Scaler	Train Error	0.0031373444682831013 in Decisior	0.0031387016960338 61 in DecisionTree
	Test Error	0.0032854942810031226 in Decisior	0.0032830414973085 32 in DecisionTree
	RMSE	1091.1126465951493 in DecisionTre	1090.2980771153843 in DecisionTree

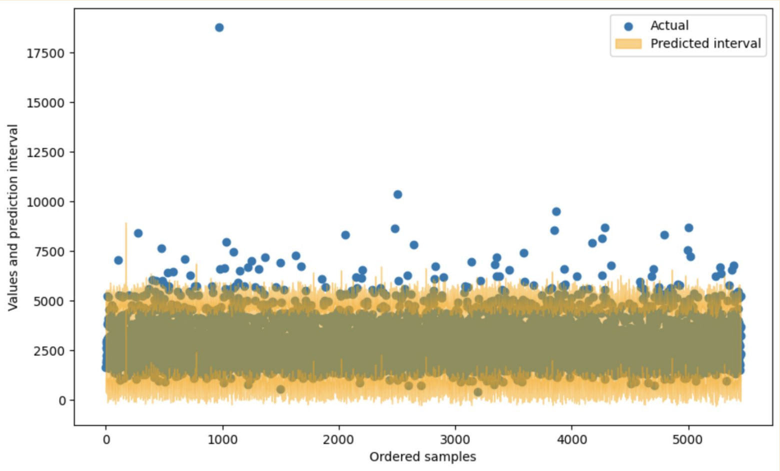
- Overall, using different scalers didn’t have a significant impact on the models’ RMSE, but we see a trend in MLP and Decision Tree yielding the least RMSE for each size

Model Selection

- **Final model:**
  - X_Train=

    estimated_store_to_consumer_driving_duration
    prep_duration_prediction
    avg_price_per_item
    estimated_order_place_duration
    price_range_of_items
    total_items

  - Y = “actual_total_delivery_duration”
  - Standard Scaler
  - MLP(Multi-Layer Perceptron) Regressor
    - Artificial Neural Network (ANN)
    - Black Box
    - Capture complex, non-linear relationships
    - Works well on datasets that are linearly separable
    - Learn interactions and adjust internal weights
- Final RMSE = 1003.11



XAI - Shapley Additive Explanations (SHAP)

- Tool for explaining the output of ML models, especially powerful for black box models

- Providing interpretability by assigning each feature a specific value that indicates its contribution to the output of the model
-
- X axis: SHAP value (feature impact on output)

- Y axis: Feature value (actual data points value for feature)

- Features ordered by their importance from high to low

XAI - Shapley Additive Explanations (SHAP)

#1: high impact with both high and low values of the feature , mix effects

#2: negative effect on the prediction for high values

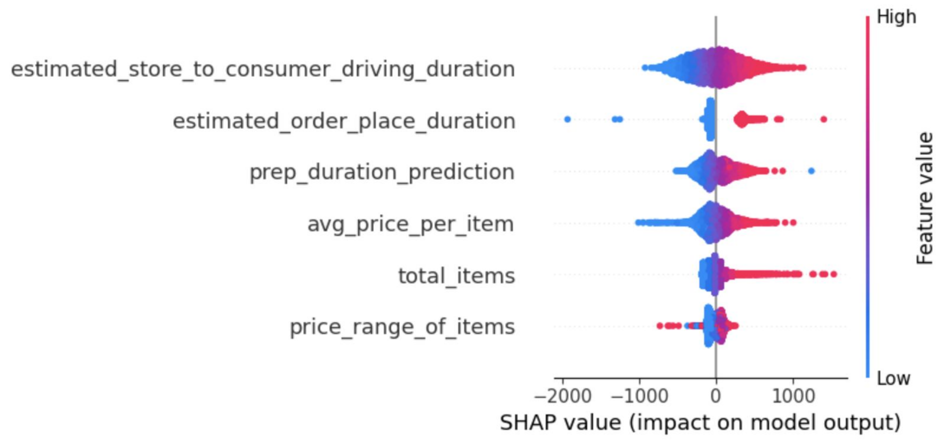
└── long order placement time ⟶ inefficiency ⟶ lower customer satisfaction

└── model trying to balance placement times & predicted outcome

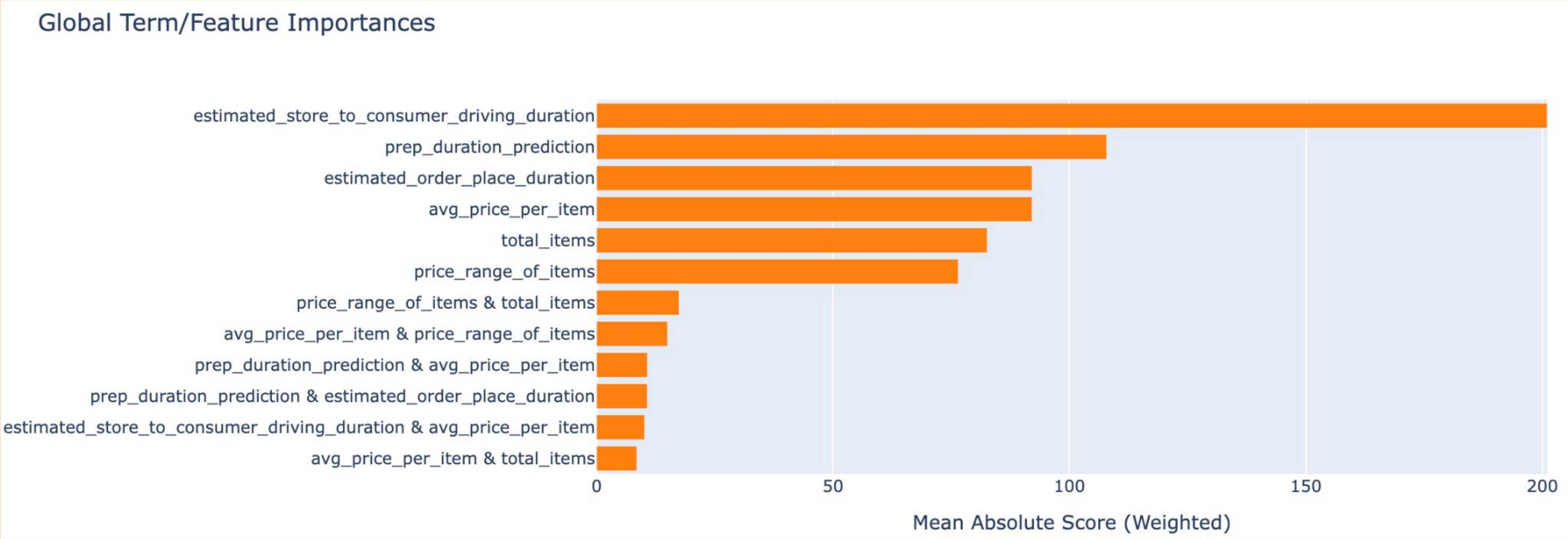
#3: mix impacts (positive > negative)

Increase: with longer driving times and more items

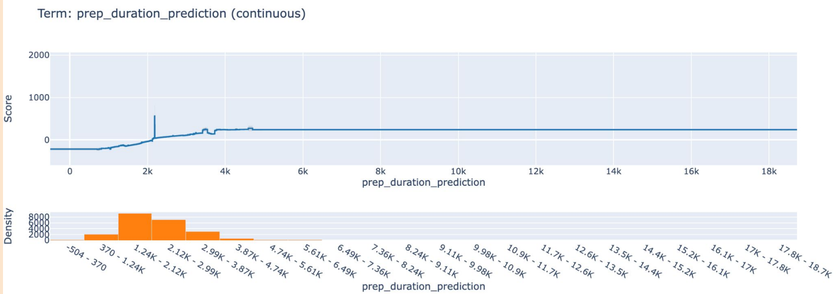
Decrease: with higher average prices per item and a smaller order size

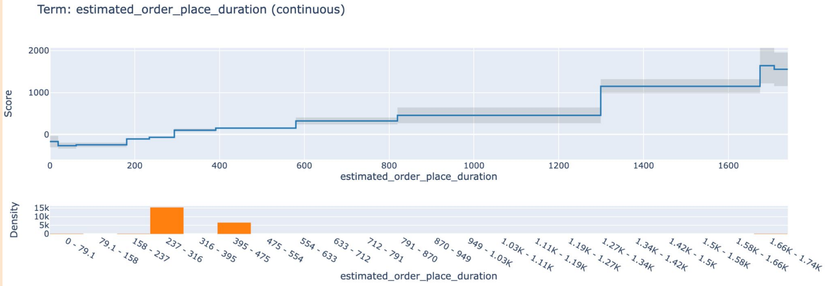


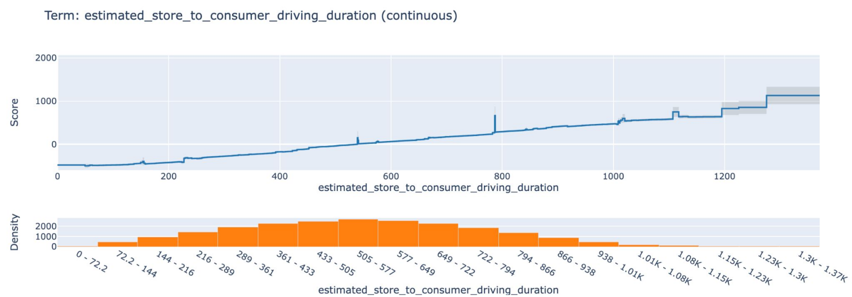
XAI - Explainable Boosting Machine (EBM)

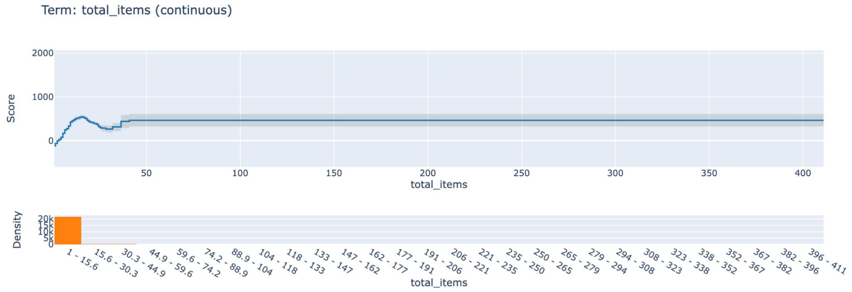


XAI - Explainable Boosting Machine (EBM)





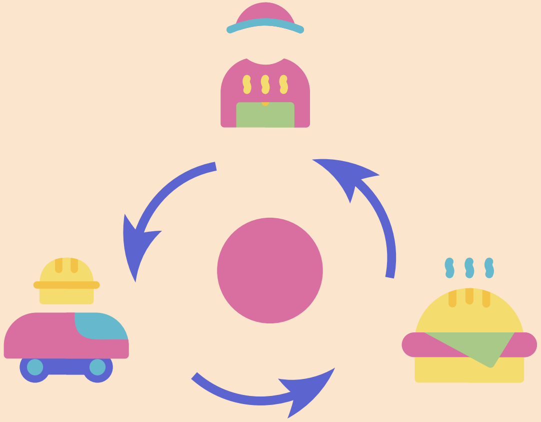




Recommendations

Customer Segmentation

Offer incentives or pricing strategies that favor orders with higher average item prices if they indeed lead to improved delivery times or customer satisfaction.



Expectation Management

Manage customer expectations and delivery logistics for orders with a large number of items, as these appear to significantly affect delivery time.

Investigation

Investigate factors that lead to efficient delivery even with longer driving durations, and replicate these conditions where possible.

Q&A

Thank You

Appendix

Feature Creation from Feature Selection Process

- Percent Distinct Item of Total = Number of Distinct Items / Total Items
- Average Price per Item = Subtotal / Total Items
- Price Range of Items = Max Item Price - Min Item Price
- Preparation Time = Actual Total Delivery Duration - Estimated Store to Consumer Driving Duration - Estimated Order Place Duration
- Preparation Duration Prediction = (Achieved via MLP Regressor using top 20 features)
- (Sum) Total Delivery Duration = Preparation Duration Prediction + Estimated Store to Consumer Driving Duration