PROJECT PRESENTATION

INSTACART MARKET BASKET ANALYSIS

Busines Problem

SETTING

A grocery ordering and delivery app, aims to make it easy to fill your refrigerator and pantry with your personal favorites and staples when you need them

CURRENT METHODS

Currently they use transactional data to develop models that predict which products a user will buy again, try for the first time, or add to their cart next during a session

RESEARCH QUESTIONS ADDRESSED

Predict whether a product will be reordered or not in the future by the customer
Predict which department will the next product ordered belong to

SOLUTION

The ability to identify which products the customers are likely to purchase again, and automatically adding those to cart through obtained predictions or provide a seamless interface for doing so will enhance their user experience

DATA EXPLORATION

MISSING VALUE IMPUTATION
OUTLIER TREATMENT
DISTRIBUTION ANALYSIS
CORRELATION TESTS
DATA VISUALIZATION
DATA PREPARATION

ORDER RELATED FEATURES

ORDER_ID

ORDER_NUMBER

AVERAGE_DAYS_BETWEEN_ORDERS

NB_ORDERS(NUMBER OF ORDERS)

TIME RELATED FEATURES

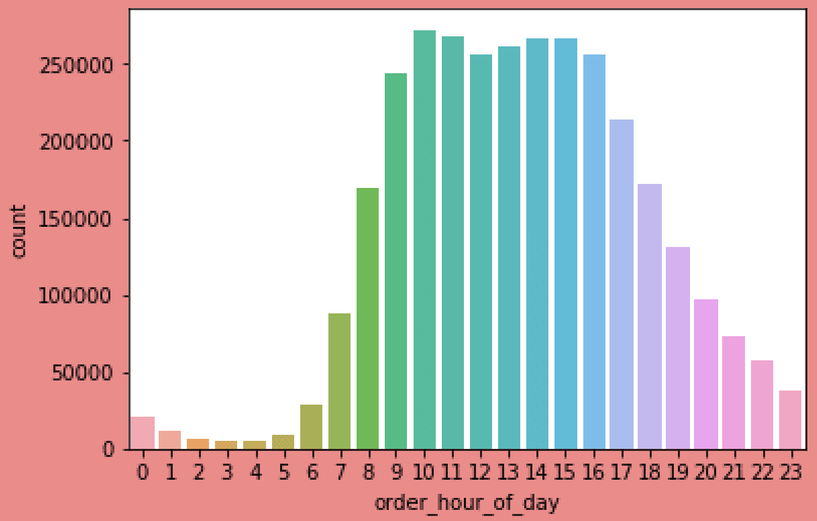
ORDER - HOUR OF THE DAY

ORDER - DAY OF THE WEED

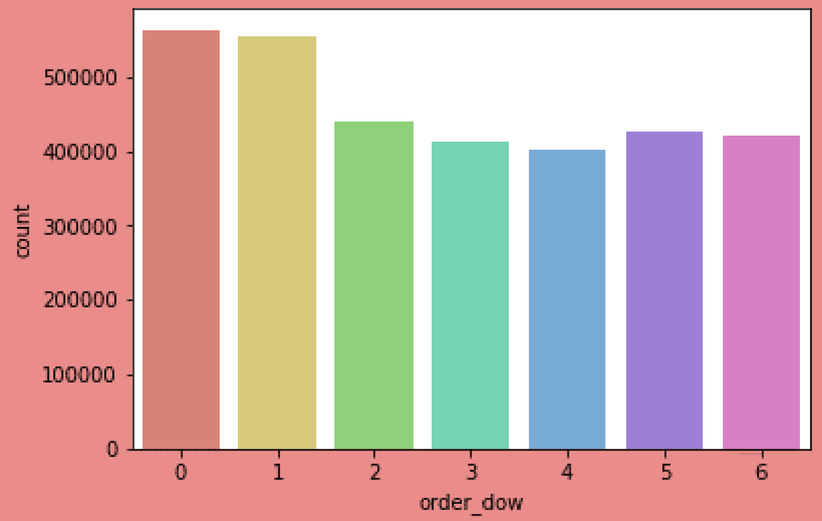
DAYS SINCE PRIOR ORDER

DAYS SINCE RATIO

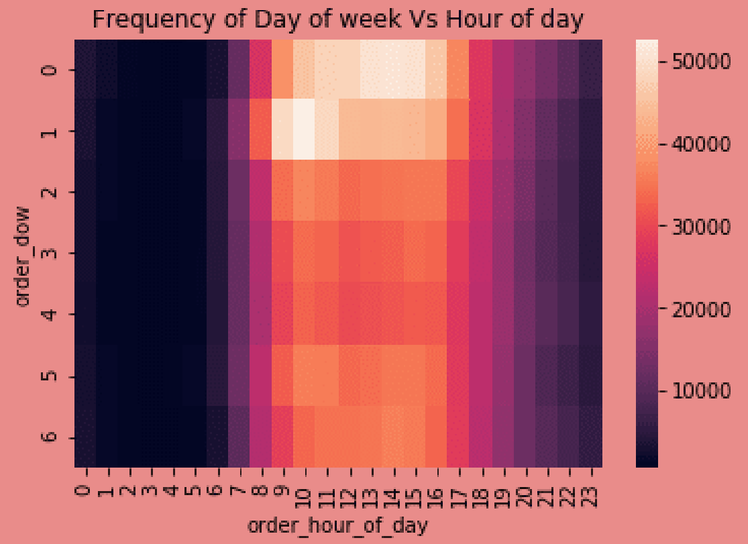
DATA FOR THE HUMAN EYE



HOUR OF THE DAY

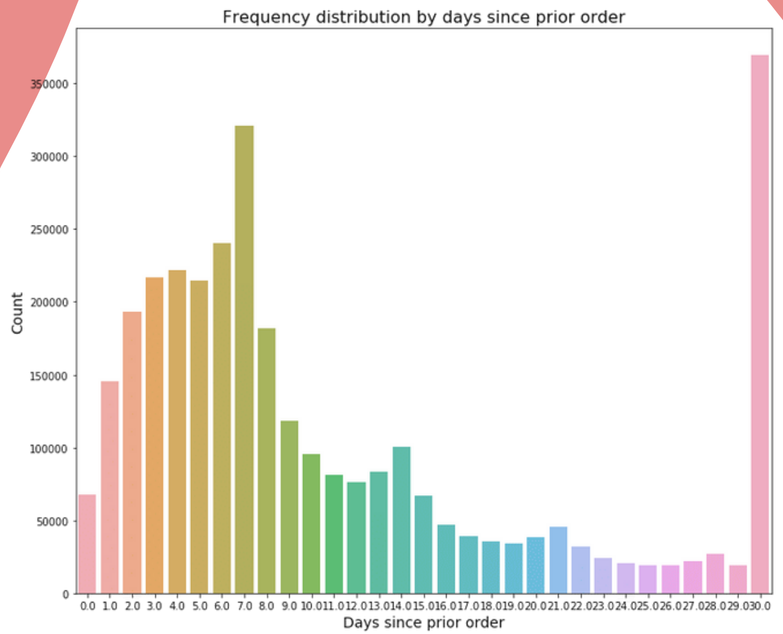


DAY OF WEEK

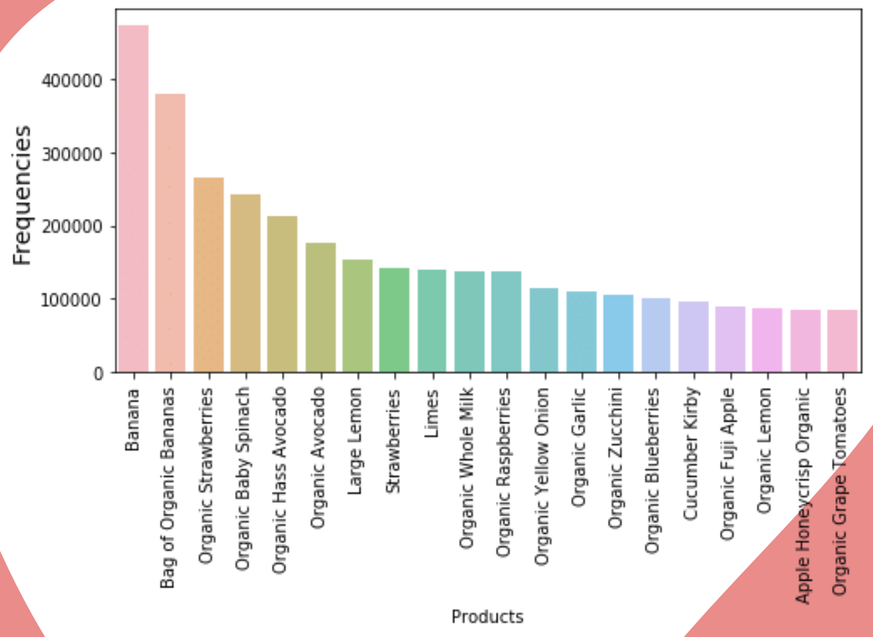


Day of week
vs
Hour of the day

DEEPER DIVE



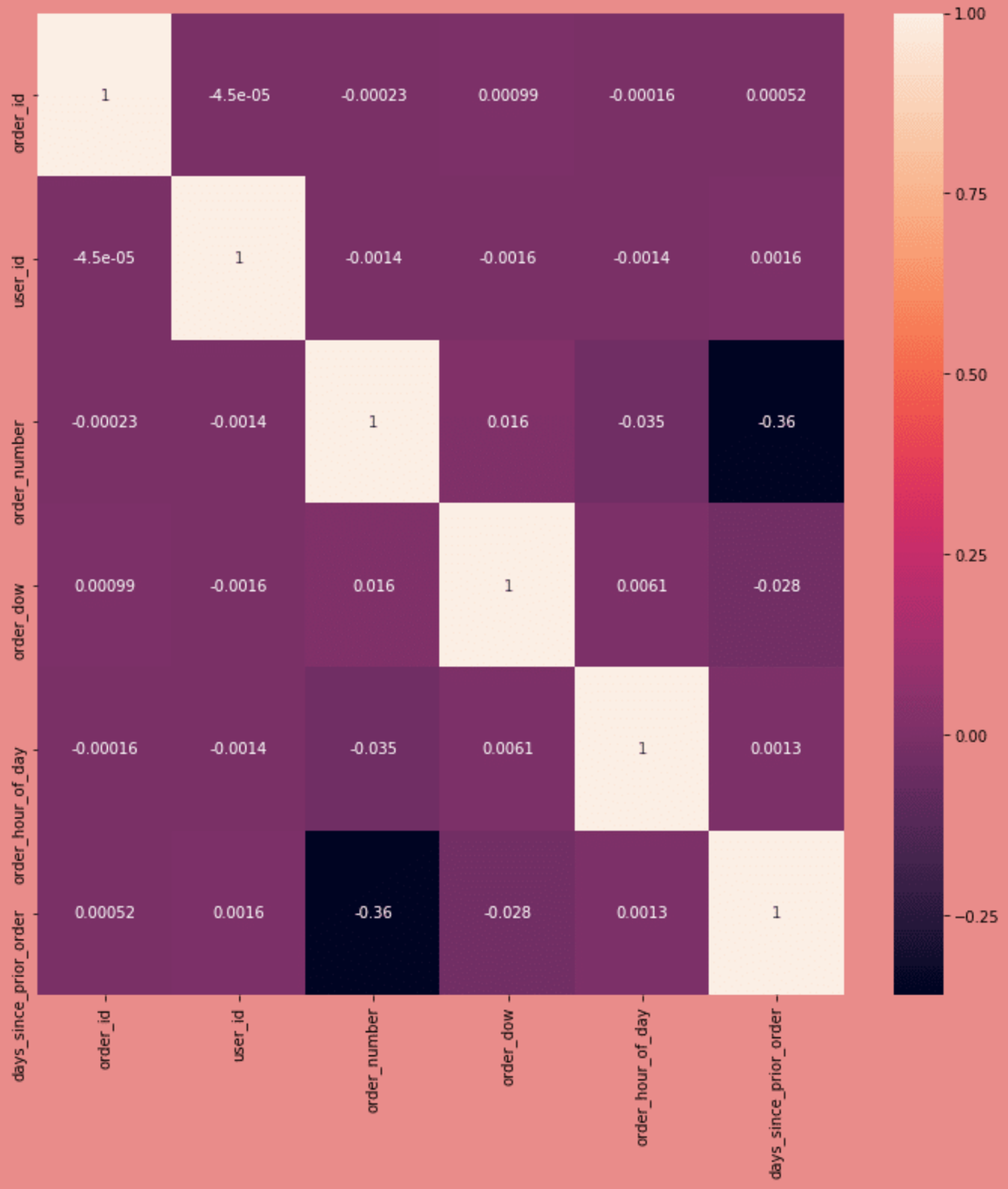
D. Since Prior Order



Top 20 products

More in the report...

CORRELATION STUDY



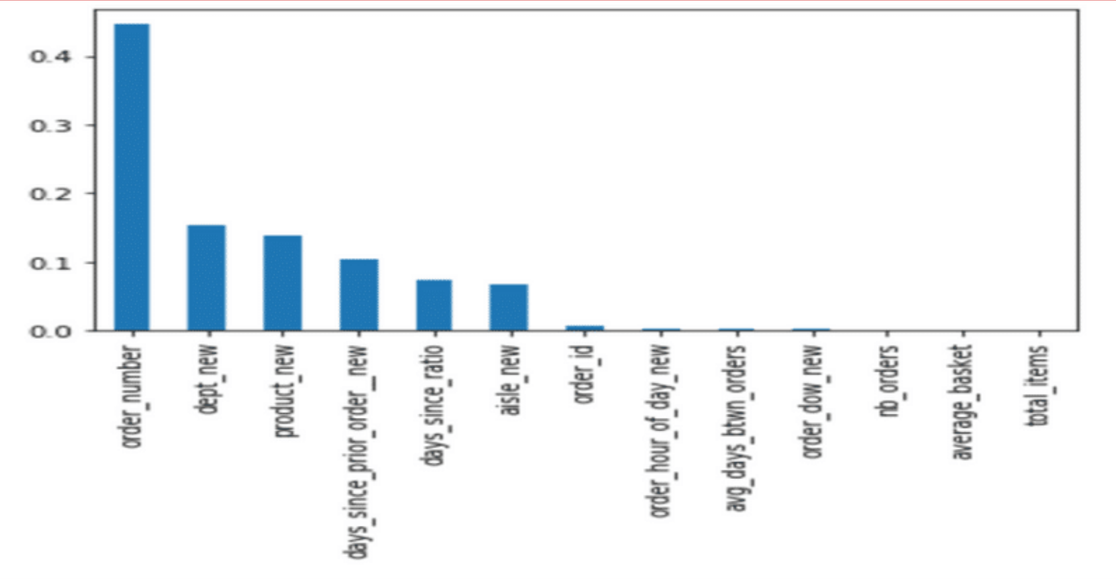
MODEL BUILDING

CHOICES

We used various models to predict if a product will be reordered or not, such as, Logistic regression, Random Forest, Adaboost Classfier and Gradient Boosting.
Logistic regression was the worst performing model while Gradient Boosting was the best performing model with an Accuracy of 67.07% on the test data. It also had the highest AUC of 0.66. Listed below are the metrics for each model on the training as well as the test data

MODEL	ACC.	PRECISION	RECALL	AUC
Logistic Regression	59.70%	85.63%	61.83%	0.56
Random Forest	59.76%	85.63%	61.83%	0.56
Adaboost Classifier	65.57%	80.02%	68.06%	0.65
Gradient Boosting	67.55%	82.87%s	69.08%	0.66

VARIABLE IMPORTANCE



- **ORDER NUMBER**
- **DEPARTMENT**
- **PRODUCT**
- DAYS SINCE PRIOR ORDER
- DAYS SINCE RATIO
- AISLE
- ORDER ID
- ORDER HOUR OF THE DAY
- AVG DAYS B/W ORDERS

MODEL 2

PREDICT DEPARTMENTAL PERFORMANCE

The objective here is to not only only say if an object belongs to one of the 21 categories, but to also provide the probability that it belongs to these classes.
We believe that the log-loss score is best suited for measuring the effectiveness of the model. Log Loss score quantifies the accuracy of the classifier by penalizing false classifications

We used Random forest, Adaboost classifier, Gradient Boosting on the dataset. Our best performing model was the Random Forest Classifier with the Lowest Log loss score of 2.342.

MODEL	LOG-LOSS SCORE
Random Forest	2.33
Adaboost Classifier	2.34
Gradient Boosting	2.97

QUESTIONS #?!

NO PRESSURE THOUGH