Decision Point Data Science Assignment

Arindam Jain
DTU
Date- 9 Oct

The question can be divided into two parts

Task 1:Predict whether Invoice will be generated or not in the next visit

Task 2: Predict the quantity of noodles that is going to orders in the next visit

Main Steps

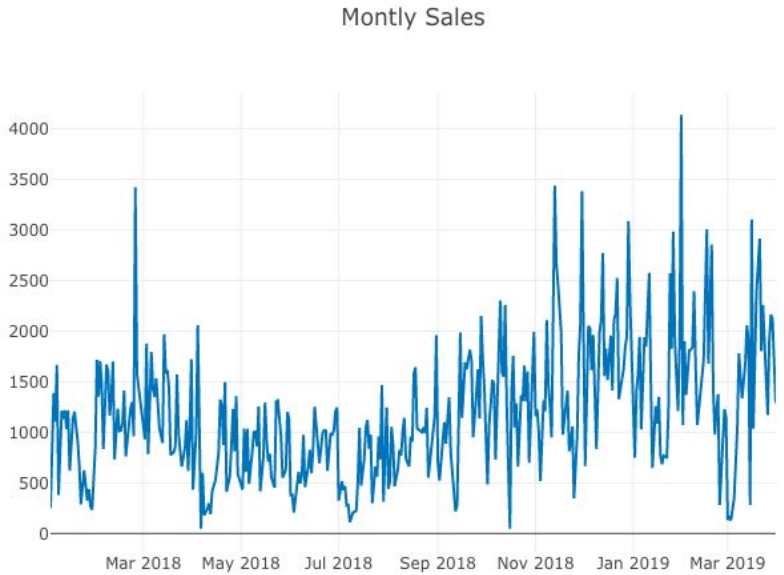
- **Step 1: Data Selection**
- **Step 2: Data Preprocessing** formatting, cleaning and sampling from it, cleaning outliers
- Step 3: **Feature Engineering** Per Unit price, Days between the last three purchases, Mean & standard deviation of the difference between purchases in days
- **Step 4: Data Transformation** Transform preprocessed data ready for machine learning by engineering features using scaling
- **Step 5: Selecting a Machine Learning Model**
- **Step 6: Evaluating the model**

Data Selection

Time Period: Jan 2018 - March 2019

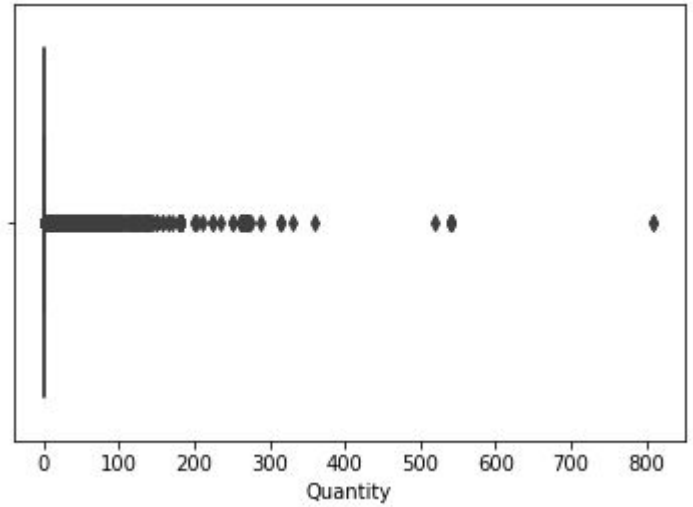
Training Data: Jan 2018 - Dec 2018

Test Data: Jan 2019 - Mar 2019

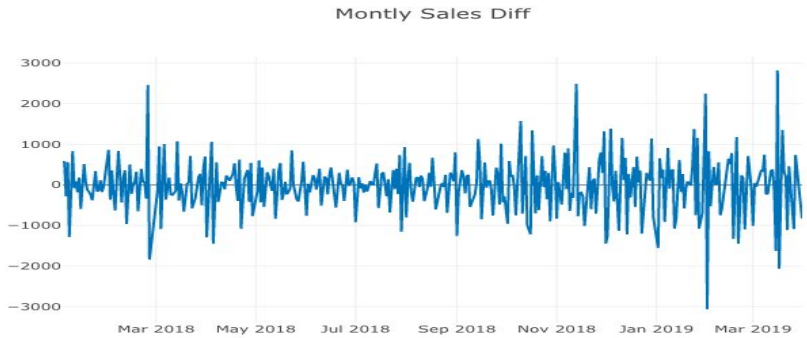


Data Preprocessing

- Check for missing values
- Check for Outliers
- Filling missing Data with mean and me
- Check for Duplicate values
- Sampling



Feature Engineering



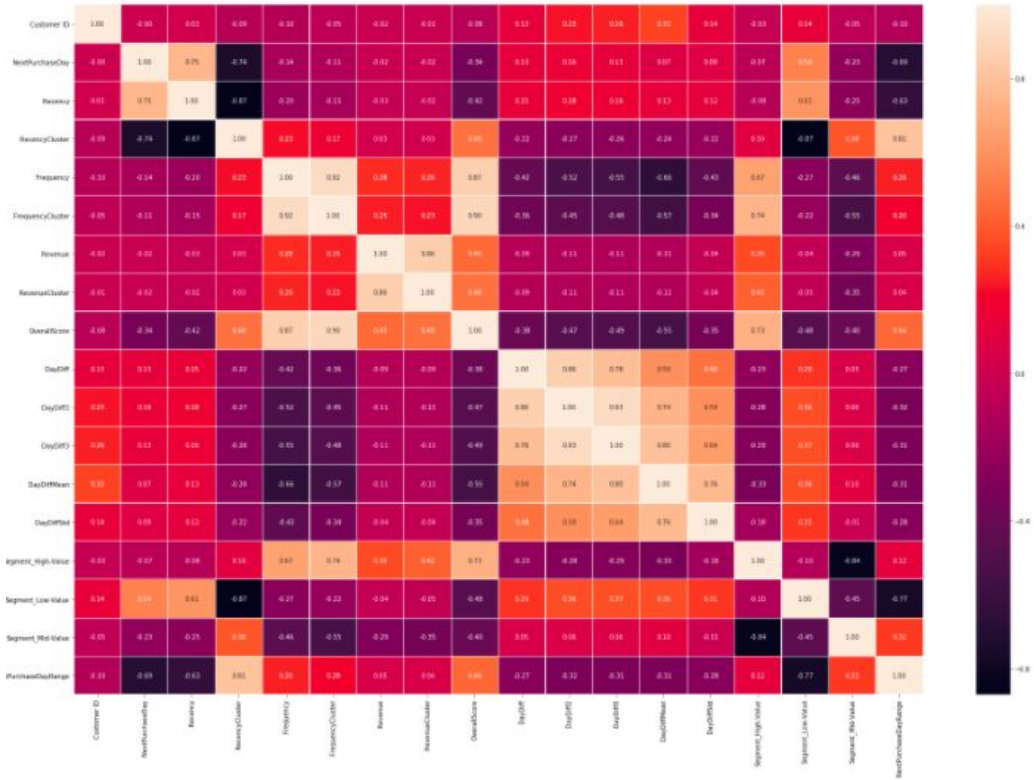
- Creating per unit price
- Days between the last three purchases

- Mean & standard deviation of the difference between purchases in days
- Experimenting with shift and lag in time series data

- To make time series stationary - log and difference in Quantity

- Scale the data

Feature Selection using correlation



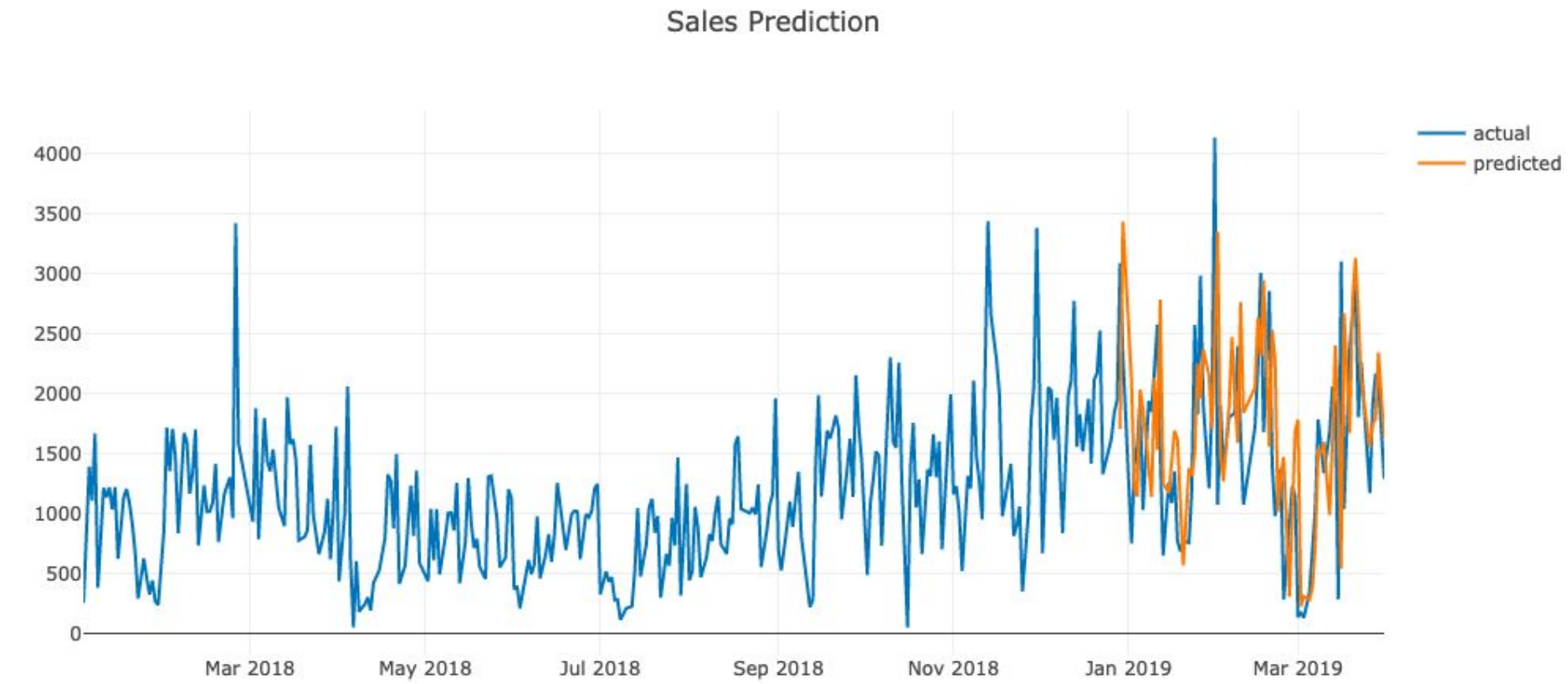
Evaluating the model

Model:	Train Accuracy	Test Accuracy
LogisticRegression.	0.91984082	0.91808874
GaussianNB	0.92154633	0.15699659
RandomForestClassifier.	0.96134167	0.96131968
SVC.	0.91984082	0.91808874
DecisionTreeClassifier.	0.93860148	0.94596132
xgb.	0.96077317	0.96075085
KNeighborsClassifier.	0.91017624	0.92093288

Xgboost gives best accuracy now will perform parameter tuning

Parameter: `'max_depth': 3, 'min_child_weight': 1`

Actual vs Predicted Sales



Performance Measurement Criteria

Precision - 0.98

Recall - 0.96

Accuracy - 0.98

F1 score - 0.97

	precision	recall	f1-score	support
0	0.82	0.95	0.88	38
1	0.38	0.70	0.49	20
2	1.00	0.97	0.98	822
avg / total	0.98	0.96	0.97	880