E-commerce

Product Recommendation and Sales Forecast

Bo Yan & Sirish Munipalli

- Problem Description
- Data Preprocessing
- Modeling Approaches and Analysis Results
  - Sales Forecast Modeling Approaches
  - Sales Forecast Analysis Results
  - Product Recommendation Modeling Approaches
  - Product Recommendation Analysis Results
- Challenges and Solutions
- Insights Gained
- Future Work

- Problem Description
- Data Preprocessing
- Modeling Approaches and Analysis Results
  - Sales Forecast Modeling Approaches
  - Sales Forecast Analysis Results
  - Product Recommendation Modeling Approaches
  - Product Recommendation Analysis Results
- Challenges and Solutions
- Insights Gained
- Future Work

Problem Description

Sales forecast:

- Sales forecast to find the potential market.
- Decision-making in overall business planning, budgeting, and risk management.
- We will be using several regression models for sales forecasting.

Product recommendations:

- Recommend products to the customer based on various product features and user purchase behaviour.
- Since the problem does not have a defined output, unsupervised learning will work perfectly for this problem.
- We will be using K-Means clustering for generating product recommendations.

- Problem Description
- Data Preprocessing
- Modeling Approaches and Analysis Results
  - Sales Forecast Modeling Approaches
  - Sales Forecast Analysis Results
  - Product Recommendation Modeling Approaches
  - Product Recommendation Analysis Results
- Challenges and Solutions
- Insights Gained
- Future Work

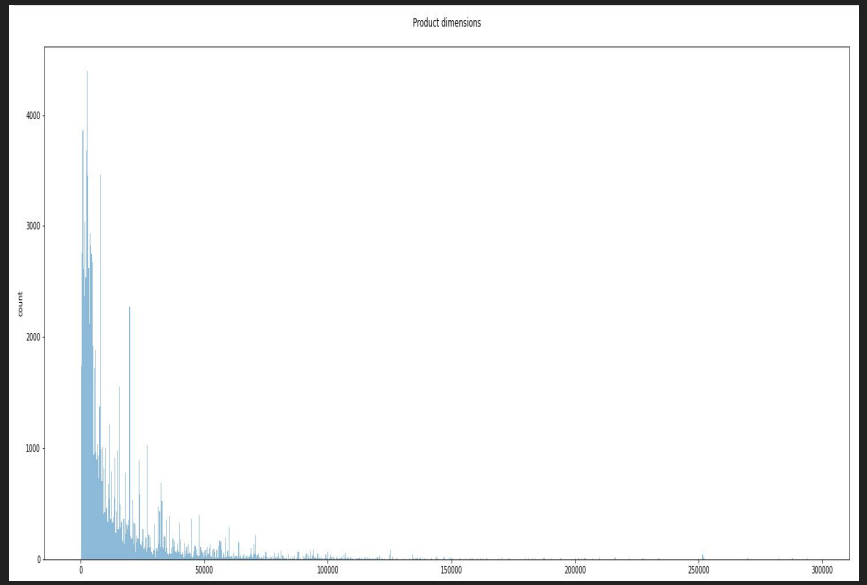
Data Preprocessing: Data cleaning

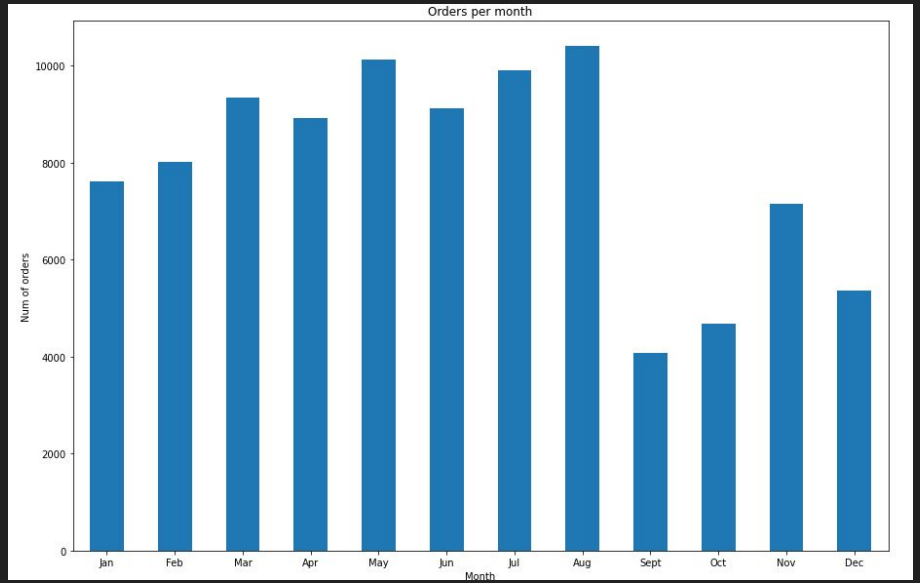
- Drop unused features
- Drop duplicated values
- Drop nullable values
- Convert data types
- Add incremental id column
- Merge datasets

Data Preprocessing: Feature Engineering

Features	Sales forecast	Product Recommendation
Price	✓	✓
Frieght_value	✓	✓
Payment_value	✓	
Survey_score	✓	✓
quarter	✓	
month		✓
product_dimension		✓
product_weight		✓
product_photo_qty		✓
product_category		✓
distance(seller-customer)		✓

Data Preprocessing: Feature visualization

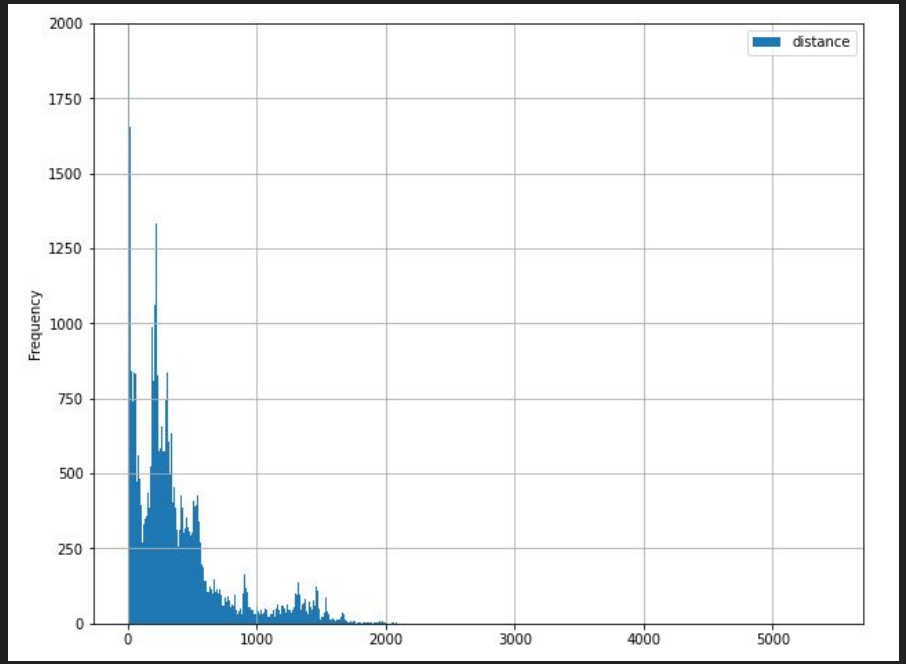




Mean = 15220.920750, std = 23264.215598
Min = 168, Max = 296208
25% = 2856, 50% = 6552
75% = 18375

March, May and Aug seem to be the months with lot of purchases.

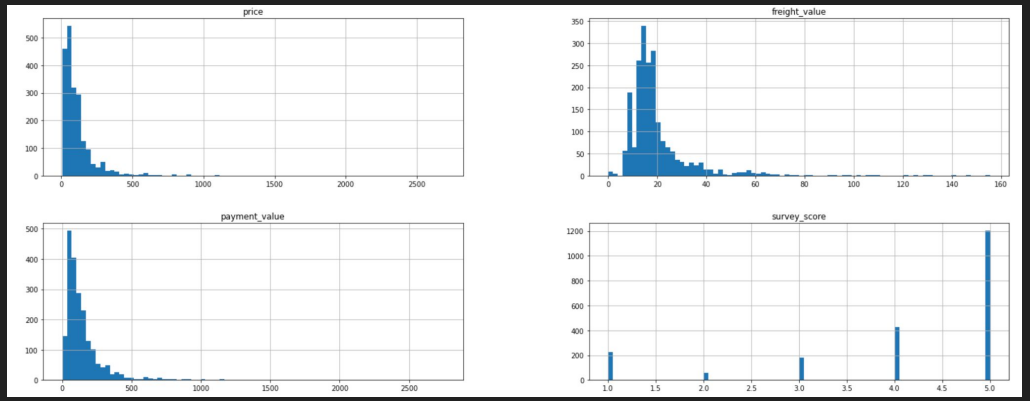
Data Preprocessing: Feature visualization



	distance
count	108113.000000
mean	370.630199
std	366.329294
min	0.000000
25%	115.236940
50%	268.571504
75%	491.492344
max	5425.108268

More than 50% of the purchases are within 300 mile radius of a seller

Data Preprocessing: Feature visualization



1. Average payment value is $153.
2. Average freight value is $20.

1. Most of the price and payment values are $100-$200.
2. Most of the survey score is 5.

	price	freight_value	payment_value	survey_score	quarter
count	2088.000000	2088.000000	2088.000000	2088.000000	2088.000000
mean	123.269090	19.822749	153.551015	4.115900	2.343870
std	182.136296	13.999481	193.904080	1.312275	1.057318
min	5.990000	0.000000	0.000000	1.000000	1.000000
25%	41.200000	13.142500	61.087500	4.000000	1.000000
50%	78.000000	16.235000	102.995000	5.000000	2.000000
75%	132.750000	20.990000	172.140000	5.000000	3.000000
max	2690.000000	155.390000	2751.240000	5.000000	4.000000

Data Preprocessing: Split datasets

Sales Forecast:

- Train: 70%
- Test: 30%

Product Recommendations:

- Train: 50%
- Test: 30%
- Validation: 20%

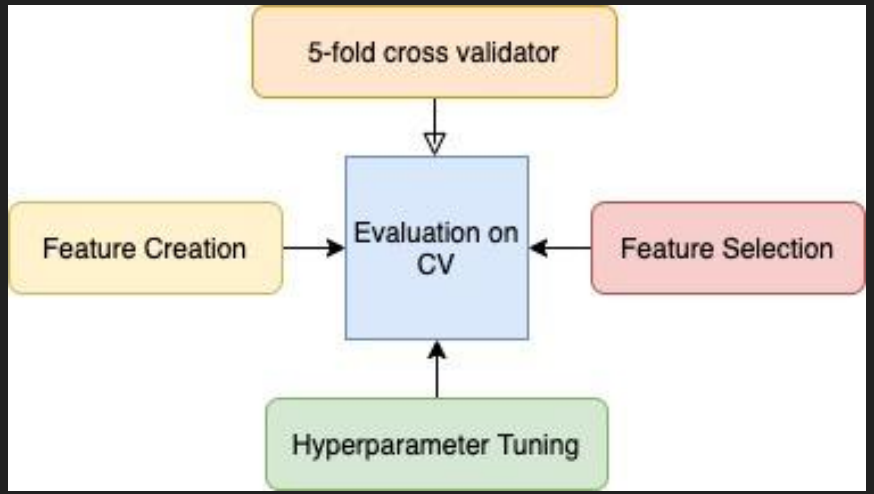
Prevent crashes:

- Limited train and test data to 30,000 and 20,000

- Problem Description
- Data Preprocessing
- Modeling Approaches and Analysis Results
  - Sales Forecast Modeling Approaches
  - Sales Forecast Analysis Results
  - Product Recommendation Modeling Approaches
  - Product Recommendation Analysis Results
- Challenges and Solutions
- Insights Gained
- Future Work

Sales Forecast Modeling Approaches

- Linear regression
  - Ridge
  - Lasso
  - Hyperparameter tuning
- Other regression models
  - Gradient-boosted tree regression
  - Decision tree regression
  - Random forest regression



Sales Forecast Modeling Approaches

Root Mean Squared Error

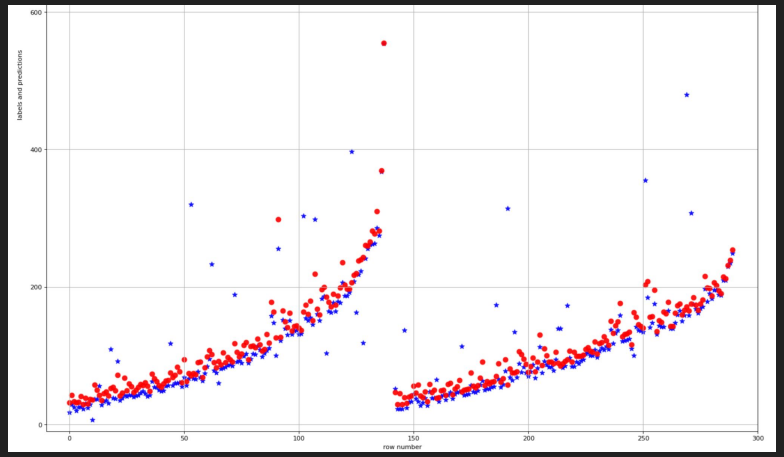
- Use RMSE to evaluate models
- Lower values of RMSE indicate better fit

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n}}$$

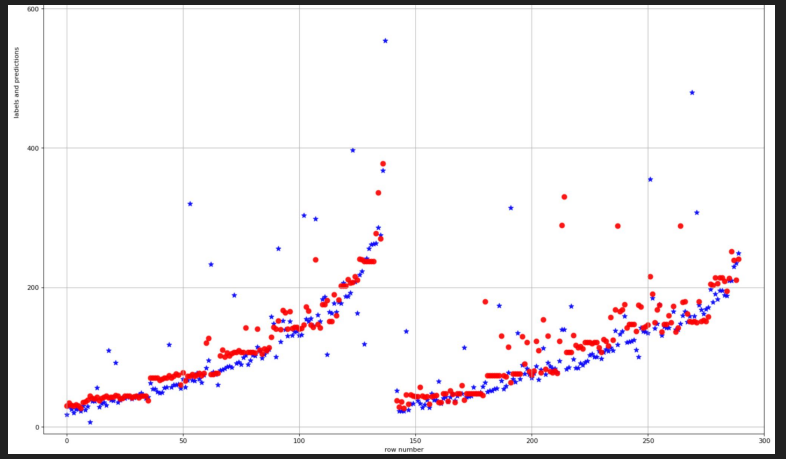
- Problem Description
- Data Preprocessing
- Modeling Approaches and Analysis Results
  - Sales Forecast Modeling Approaches
  - Sales Forecast Analysis Results
  - Product Recommendation Modeling Approaches
  - Product Recommendation Analysis Results
- Challenges and Solutions
- Insights Gained
- Future Work

Sales Forecast Analysis Results

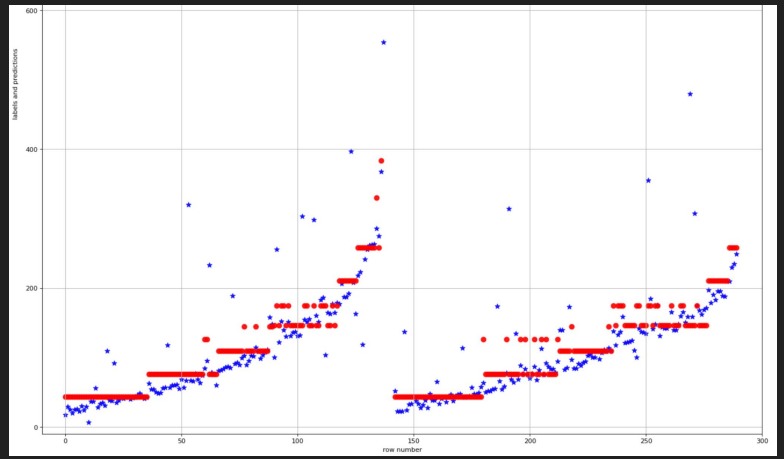
Linear Regression



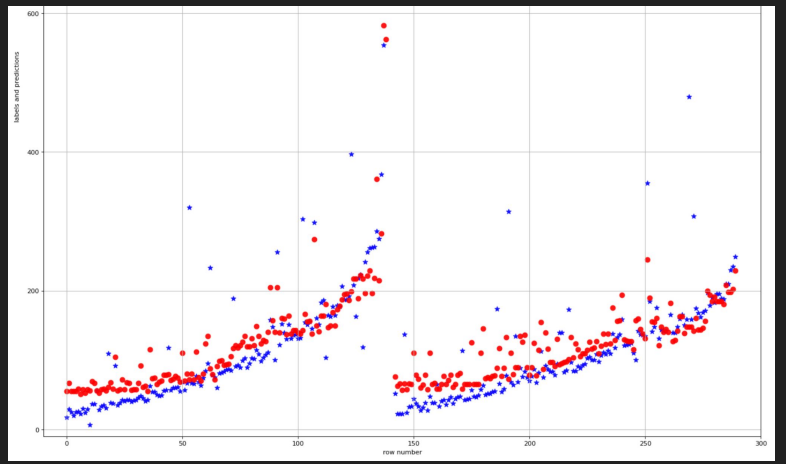
Gradient-boosted tree regression



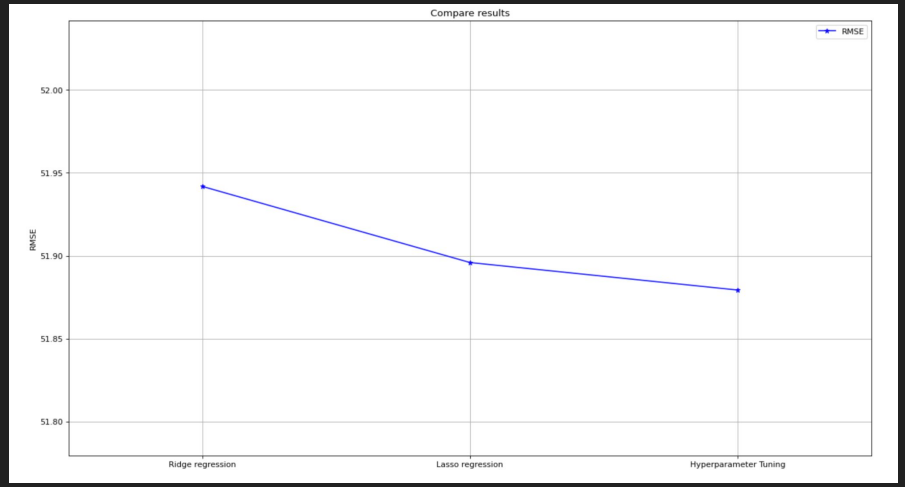
Decision Tree Regression



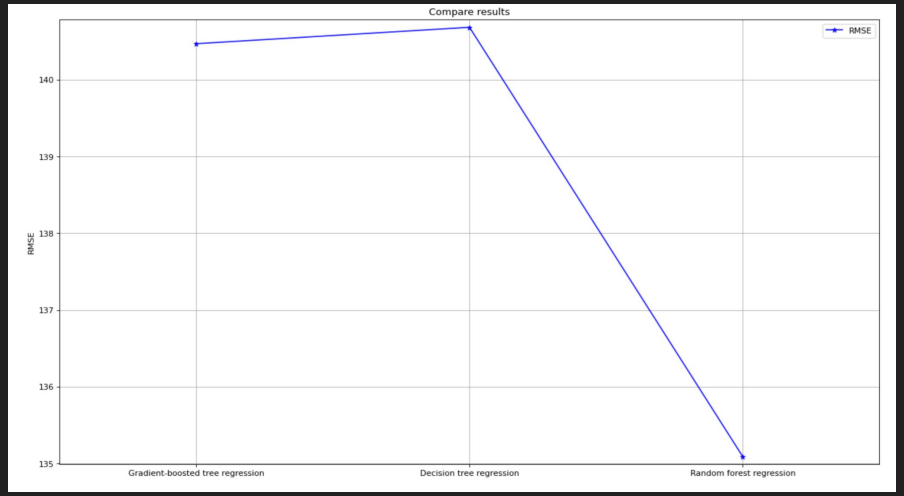
Random forest regression



Sales Forecast Analysis Results



The linear regression performs better after hyperparameter tuning.



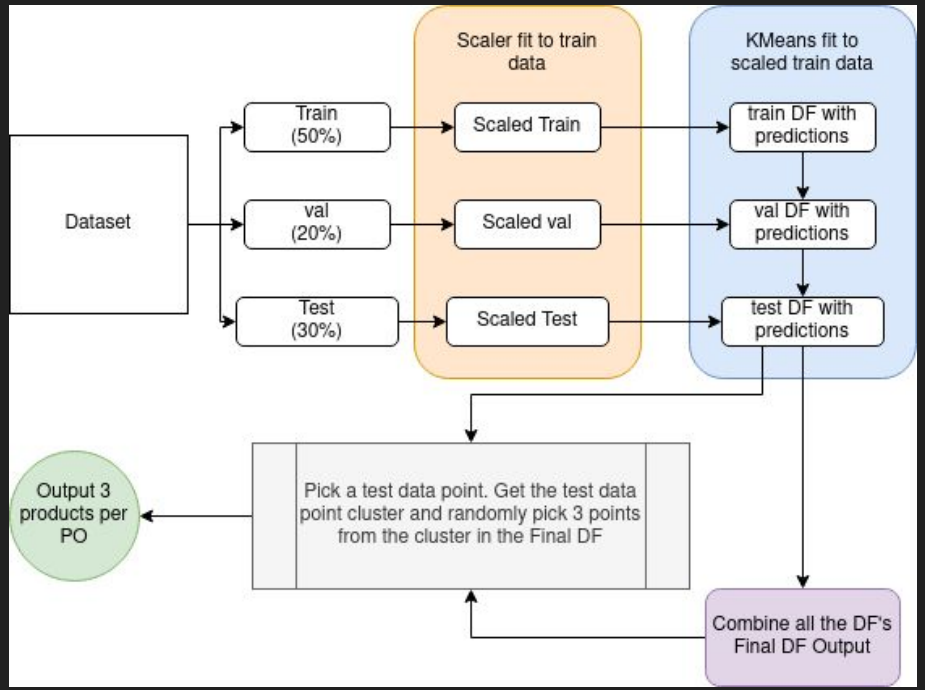
The random forest regression performs better than Gradient-boosted tree regression and Decision tree regression.

- Problem Description
- Data Preprocessing
- Modeling Approaches and Analysis Results
  - Sales Forecast Modeling Approaches
  - Sales Forecast Analysis Results
  - Product Recommendation Modeling Approaches
  - Product Recommendation Analysis Results
- Challenges and Solutions
- Insights Gained
- Future Work

Product Recommendation Modeling Approaches

Product recommendations:

- K-means clustering to create the clusters. By using the features mentioned in the features slide.
- All the output DF’s contain the product_id, customer_id and prediction columns



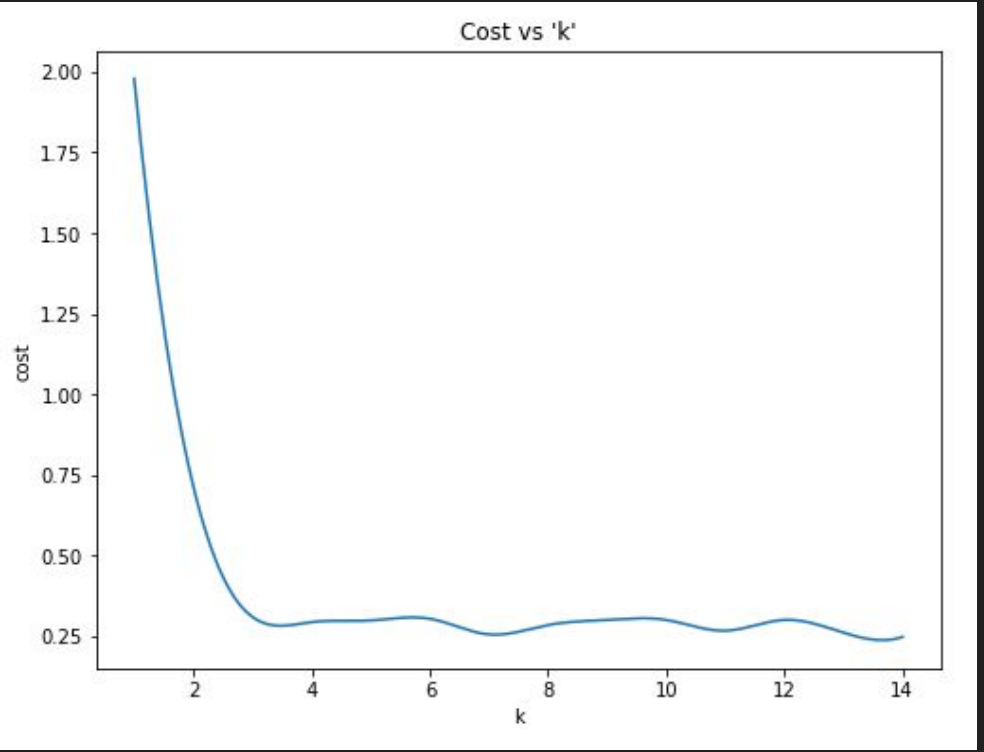
Product Recommendation Modeling Approaches

Optimal ‘k’ value:

- Using Silhouette score to get the optimal number of clusters

$$Silhouette - score = \frac{b_i - a_i}{\max(bi, a_i)}$$

- After multiple runs a cluster value of 4 seems to do a good job.

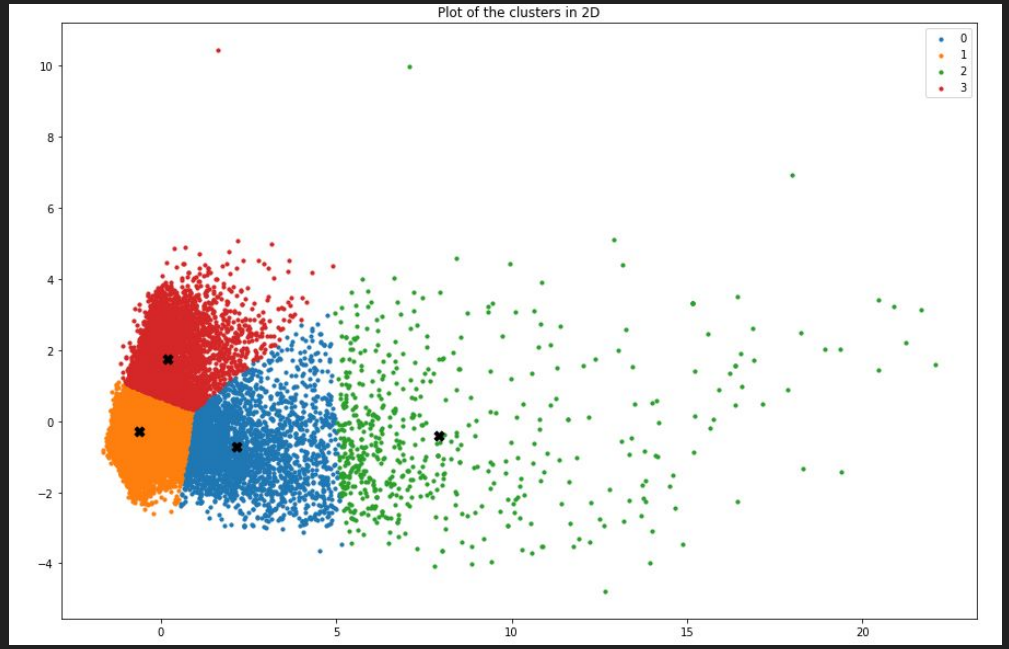


- Problem Description
- Data Preprocessing
- Modeling Approaches and Analysis Results
  - Sales Forecast Modeling Approaches
  - Sales Forecast Analysis Results
  - Product Recommendation Modeling Approaches
  - Product Recommendation Analysis Results
- Challenges and Solutions
- Insights Gained
- Future Work

Product Recommendation Analysis Results

- Using PCA to reduce the 9 feature dimensions to 2 for plotting.
- As it can be seen that 3 clusters are very distinct but the 4 cluster is very spares.
- A better feature scaling may fix this issue or we might need more data.
-

Run	Train	Validation	Test
Scores	0.294	0.2928	0.298



Product Recommendation Analysis Results

Product : 6b75ce117b8fcc75289cb6cbe589de6c purchased by the customer :71af36f53c0fe0b5b886ffad8154c5db
Top 3 recommended products for the specific customer

distance	survey_score	month	price	freight_value	product_dim_cm	product_weight_g	product_photos_qty	cat_id	product_id	customer_id	features	scaled	prediction
6.478096287460243	5	4	117.3	12.81	44890	4105	1	2	a50acd33ba7a8da8e...	2c94ee4423f153e13...	[6.4780962874602 4...	[0.01774078494407...	1
12.265322028224537	4	4	219.0	9.3	2700	292	2	9	d04e48982547095af...	63368406d4bb7f949...	[12.265322028224 5...	[0.03358956562498...	1
30.60662347612332	5	12	79.99	8.75	8960	250	1	22	ccc4bd1600ccbb62e...	a0edaa97421af6b1d...	[30.606623476123 3...	[0.08381868698144...	1

Product : a2da86fa759178e9e58e54aa1a144e59 purchased by the customer :68cb7fbc85416655ad0499fcc7fdb9f7
Top 3 recommended products for the specific customer

distance	survey_score	month	price	freight_value	product_dim_cm	product_weight_g	product_photos_qty	cat_id	product_id	customer_id	features	scaled	prediction
6.271393215280431	5	4	29.9	7.39	15840	150	6	0	adf591c625cb265c1...	47ed1ad73b7883c3f...	[6.2713932152804 3...	[0.01717471204424...	2
7.028794344256277	5	8	30.0	8.37	11270	500	6	43	37f4d0bf85fbf875c...	dc7b3fbf491155c69...	[7.028794344256 27...	[0.01924891562957...	2
14.914633111221498	5	1	134.17	27.38	43560	10400	10	11	9523f1a3e7db9e38d...	a8c5c582c453d8fe3...	[14.914633111221 4...	[0.04084491597604...	2

Product : dfb97c88e066dc22165f31648efe1312 purchased by the customer :2c94ee4423f153e13ce3fb15ac406a13
Top 3 recommended products for the specific customer

distance	survey_score	month	price	freight_value	product_dim_cm	product_weight_g	product_photos_qty	cat_id	product_id	customer_id	features	scaled	prediction
18.645394688770942	5	1	49.9	8.27	8000	100	2	6	473795a355d29305c...	b930ac822ab2d45dd...	[18.645394688770 9...	[0.05106190502466...	1
26.103049476398382	5	3	110.0	13.05	50400	7800	3	44	88e84a987b4681434...	7302aa13024a3e490...	[26.103049476398 3...	[0.07148528928812...	1
26.528569471573803	4	1	155.0	8.83	5967	350	3	33	e40fa115d27ea4a80...	7e27f4fc651dbbb8e...	[26.528569471573 8...	[0.07265060983737...	1

Product : 37f4d0bf85fbf875c920d460766d6a5c purchased by the customer :db7432cb997db7083db6aaea715d3433
Top 3 recommended products for the specific customer

distance	survey_score	month	price	freight_value	product_dim_cm	product_weight_g	product_photos_qty	cat_id	product_id	customer_id	features	scaled	prediction
22.111542860294147	5	3	49.9	10.96	3136	600	4	7	1720f85e4a07f5701...	e785a41a19890a8f1...	[22.111542860294 1...	[0.06055422909128...	2
24.10704168477952	5	12	118.6	8.09	2288	250	7	17	7a7416293b9692d84...	9f11f34a505da5700...	[24.107041684779 5...	[0.06601906226609...	2
25.923252137933993	4	1	129.9	9.1	4800	550	4	23	f0f464c1300173b9e...	0b79f3d959c1308c5...	[25.923252137933 9...	[0.07099289989258...	2

Product : 2136c70bbe723d338fab53da3c03e6dc purchased by the customer :70a8cfb1730fd53e5c15f2a62e1e5448
Top 3 recommended products for the specific customer

distance	survey_score	month	price	freight_value	product_dim_cm	product_weight_g	product_photos_qty	cat_id	product_id	customer_id	features	scaled	prediction
13.19973867236129	1	8	5.9	7.39	4590	250	1	7	9a4b54310a69c82a6...	16d5e579f5c682f59...	[13.199738672361 2...	[0.03614854036018...	0
16.74617105631575	1	5	579.99	16.96	16000	3750	3	25	71f2ce4ae5dfdd3b5...	ee78e2394f33a3bf7...	[16.746171056315 7...	[0.04586072916543...	0
19.282037475200585	1	5	38.7	8.29	4096	450	1	29	c9e5a053551073d27...	705b9e5a5d27aaed4...	[19.282037475200 5...	[0.05280540222802...	0

- Problem Description
- Data Preprocessing
- Modeling Approaches and Analysis Results
  - Sales Forecast Modeling Approaches
  - Sales Forecast Analysis Results
  - Product Recommendation Modeling Approaches
  - Product Recommendation Analysis Results
- Challenges and Solutions
- Insights Gained
- Future Work

Challenges and Solutions

1. Crashes due to large data size and lot of features.
    a. A better hardware and a cloud computing resource can fix this issue.
2. Feature scaling: Outliers in the output
    a. More work needs to be done in feature scaling
3. Product picking Scheme: In the current implementation we are randomly picking a product from the test_output cluster. Sometimes might now get us the desired product in the recommendations
    a. Use a distance metric to choose the recommended products.
    b. Implement another layer at the end for choosing a product.
4. Lack of needed features, such as product quantity, discount values.
    a. Use order payments values and quarter as selected features.

- Problem Description
- Data Preprocessing
- Modeling Approaches and Analysis Results
  - Sales Forecast Modeling Approaches
  - Sales Forecast Analysis Results
  - Product Recommendation Modeling Approaches
  - Product Recommendation Analysis Results
- Challenges and Solutions
- Insights Gained
- Future Work

Insights Gained

Sales forecast:

- Hyperparameter tuning helps choose a set of optimal hyperparameters which performs better in reducing RMSE for E-commerce data.
- Linear regression performs better than other three regressions for E-commerce data.

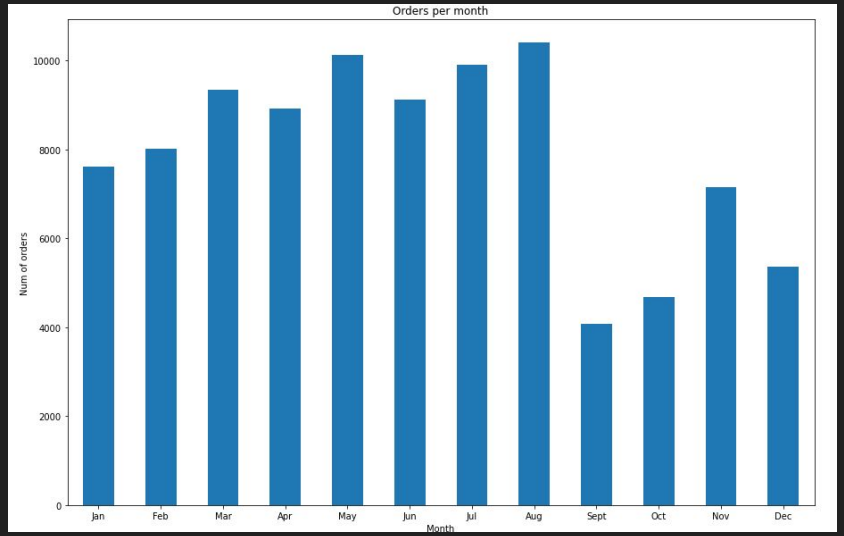
Product Recommendations:

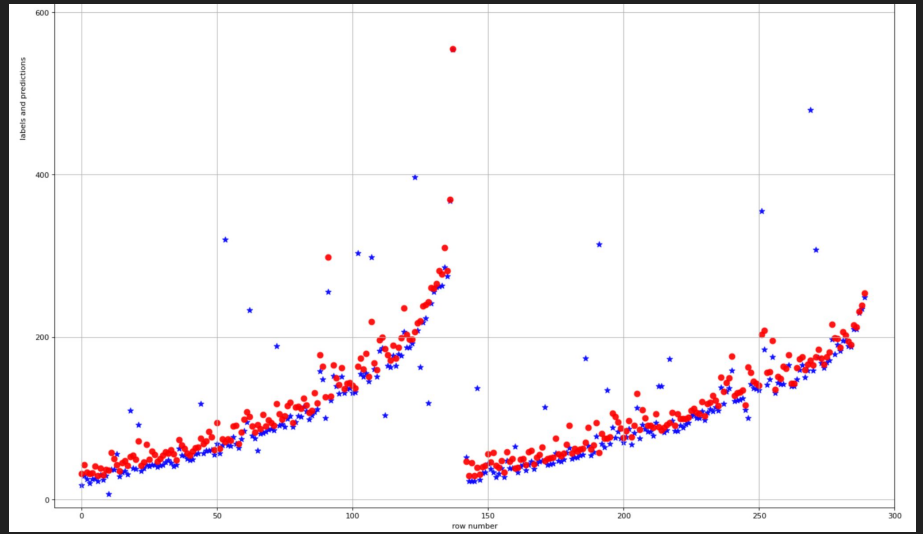
- Explore other unsupervised algorithms and SVD
- Take other customers buying patterns into account
- Use a better hardware or cloud computing platform
- Work on a better feature scaling and more data

- Problem Description
- Data Preprocessing
- Modeling Approaches and Analysis Results
  - Sales Forecast Modeling Approaches
  - Sales Forecast Analysis Results
  - Product Recommendation Modeling Approaches
  - Product Recommendation Analysis Results
- Challenges and Solutions
- Insights Gained
- Future Work

Future Work

- Since the data has seasonal patterns, we will use time series models to predict sales, like AR, MA, ARMA, ARIMA, and ES models.
- There might be other factors that have effects on the time series data that we need to consider, for example, seasonal variations, irregular variations.





Thank you!

Q&A