Capstone Project

Retail Sales Prediction

Midhun R

Points for Discussion

- Business Task
- Data Summary
- Data Cleaning
- Exploratory Data Analysis
- Feature Engineering
- Modelling
- Conclusion

Business Task

Two datasets are given: one with store data of 1115 stores and the other with historical sales data from January 2013 to July 2015.

The main objective is to analyze the given dataset and build a machine learning model to forecast the sales of all Rossmann stores upto 6 weeks.

This is undertaken as an individual project.

Data Summary

	Store Data	Sales Data
Number of records (rows)	1115	1017209
Number of features (columns)	10	9
Number of duplicate rows	0	0
Number of columns with missing values	6	0
Number of columns require conversion of data type	4	2

- In addition to this, extra columns are needed to be added for easier analysis.
- These irregularities will be handled during data cleaning step and the two datasets will be merged.

Data Summary (Contd.)

Store Data:

1. Store: a unique Id for each store.
2. StoreType: differentiates between 4 different store models: a, b, c, d.
3. Assortment: describes an assortment level: a = basic, b = extra, c = extended.
4. CompetitionDistance: distance in meters to the nearest competitor store.
5. CompetitionOpenSinceMonth: gives the approximate month of the time the nearest competitor was opened.
6. CompetitionOpenSinceYear: gives the approximate year of the time the nearest competitor was opened.
7. Promo2: Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating.
8. Promo2SinceWeek: describes the calendar week when the store started participating in Promo2.
9. Promo2SinceYear: describes the year when the store started participating in Promo2.
10. Promo2Interval: describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew. Eg. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store.

Data Summary (Contd.)

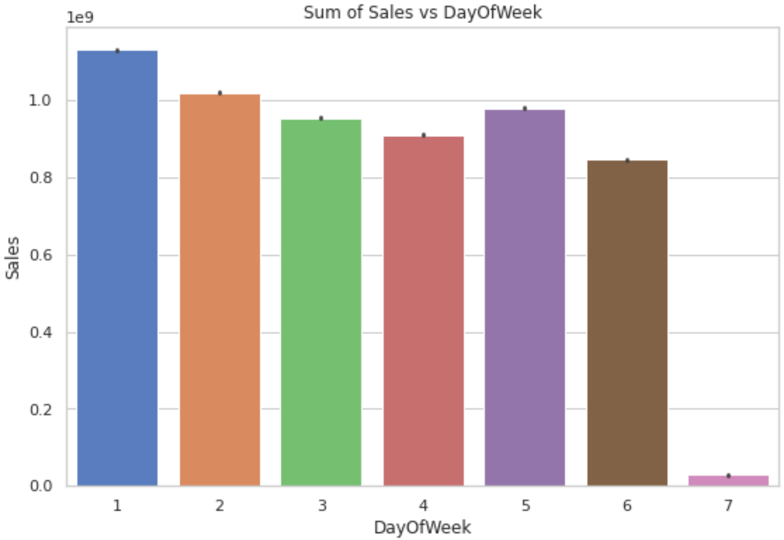
Sales Data:

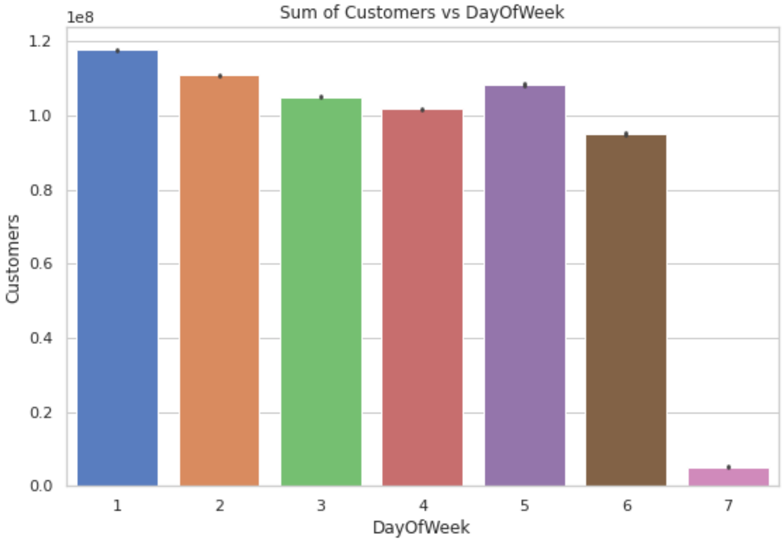
1. Store: a unique Id for each store.
2. DayOfWeek: day of week of sale.
3. Date: date of sale
4. Sales: the turnover for any given day.
5. Customers: the number of customers on a given day.
6. Open: an indicator for whether the store was open: 0 = closed, 1 = open.
7. Promo: indicates whether a store is running a promo on that day.
8. StateHoliday: indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None.
9. SchoolHoliday: indicates if the (Store, Date) was affected by the closure of public schools.

Data Cleaning

- 3 rows of data in ‘CompetitionDistance’ were null, it was imputed with the median of data.
- About 32% of data in ‘CompetitionOpenSinceMonth’ & ‘CompetitionOpenSinceWeek’ was null, they were imputed with the mode of data.
- About 49% of data in ‘Promo2SinceWeek’, ‘Promo2SinceYear’ & ‘PromoInterval’ was null in the same rows, where the value of Promo2 is zero. So, the missing values were imputed with 0.
- After the merging of two datasets, ‘Date’ was converted to datetime datatype and ‘CompetitionOpenSinceMonth’, ‘CompetitionOpenSinceYear’, ‘Promo2SinceWeek’ & ‘Promo2SinceYear’ to int datatype.
- ‘StateHoliday’ was converted into a binary feature.
- Five new features ‘WeekOfYear’, ‘Month’, ‘Year’, ‘CompetitionOpenNumMonths’ and ‘Promo2NumWeeks’ were added.

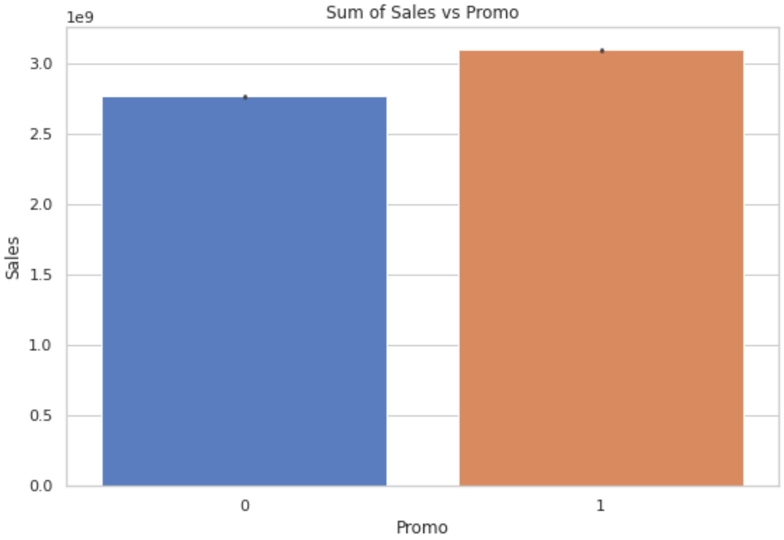
Exploratory Data Analysis

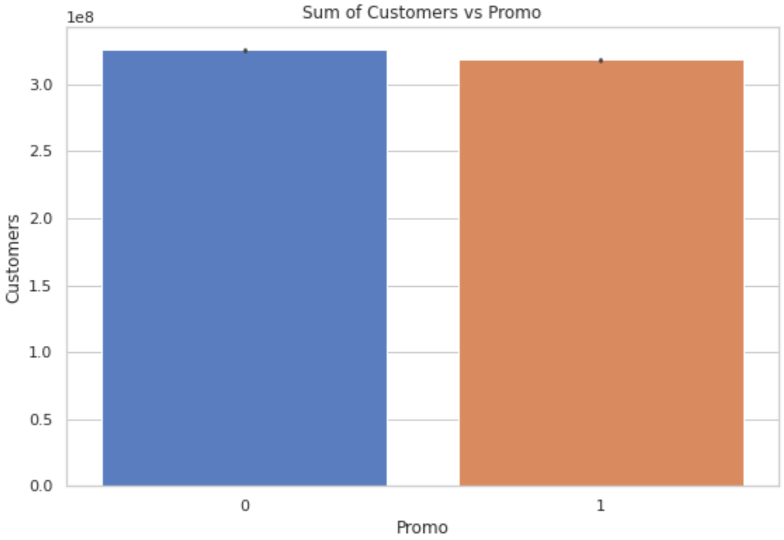




Highest sales are recorded on Mondays and lowest sales are recorded on Sundays. This may be because most of the shops are closed on Sundays and this leads to higher demand on the next day, which is Monday.

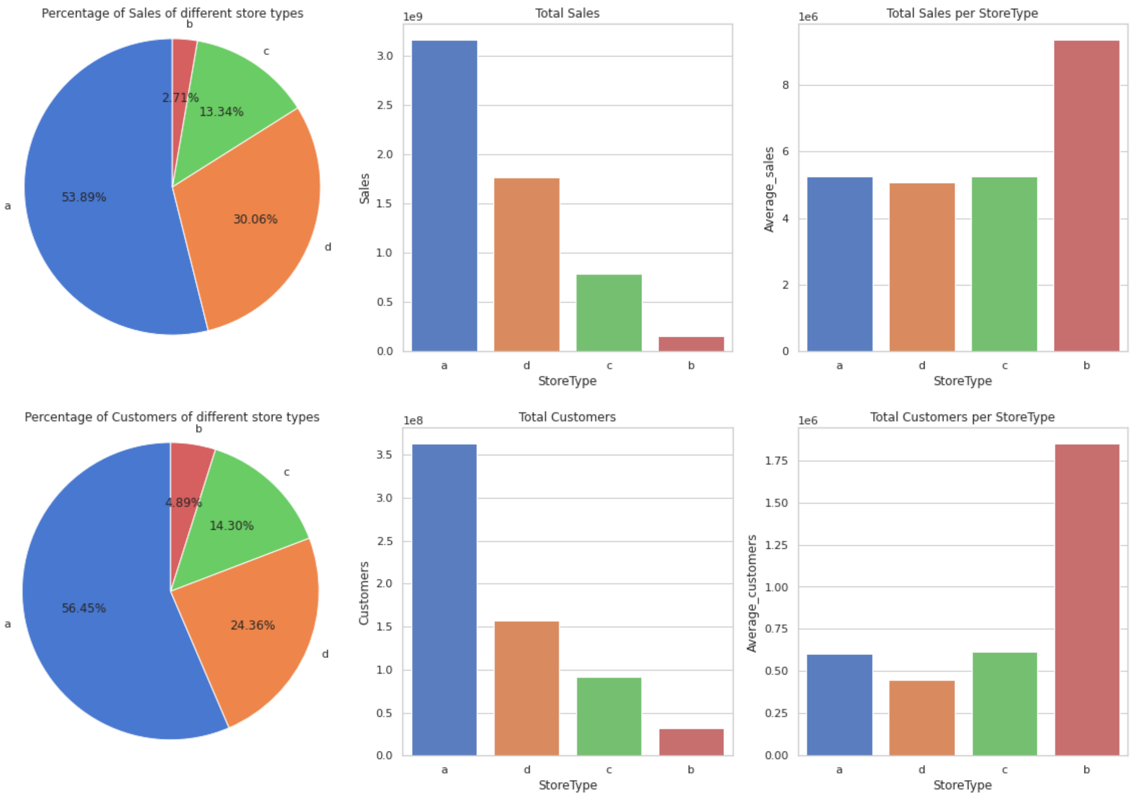
Exploratory Data Analysis (Contd.)





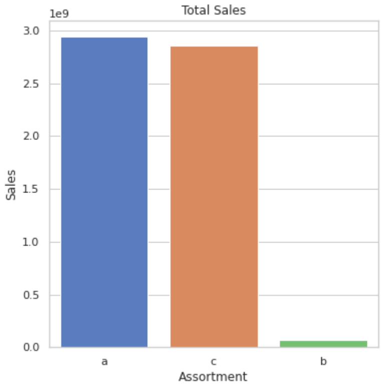
Presence of promos increases sales, but it doesn't help much in generating new customers.

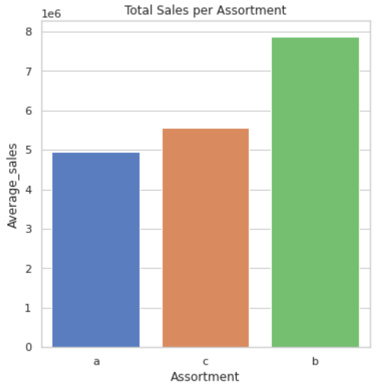
Exploratory Data Analysis (Contd.)

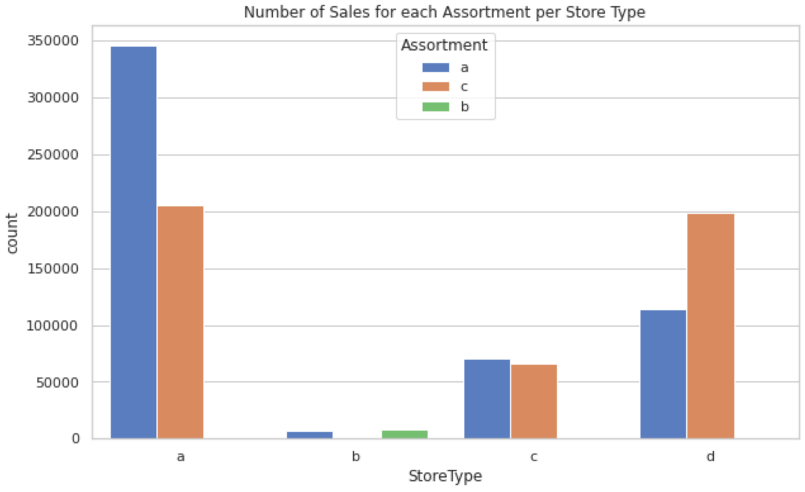


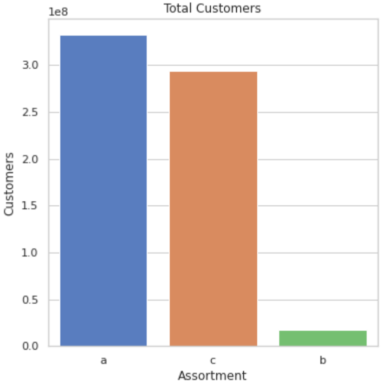
Even though the volume of sales and customers is low, store type b has the highest average sales and customers. This means store type b is more preferred by customers.

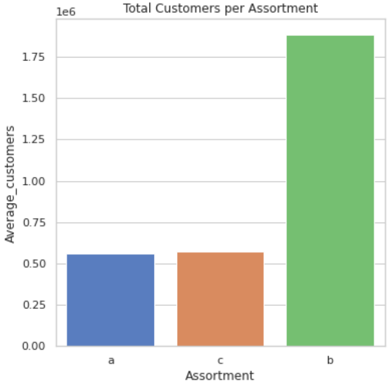
Exploratory Data Analysis (Contd.)


Total Sales


Total Sales per Assortment

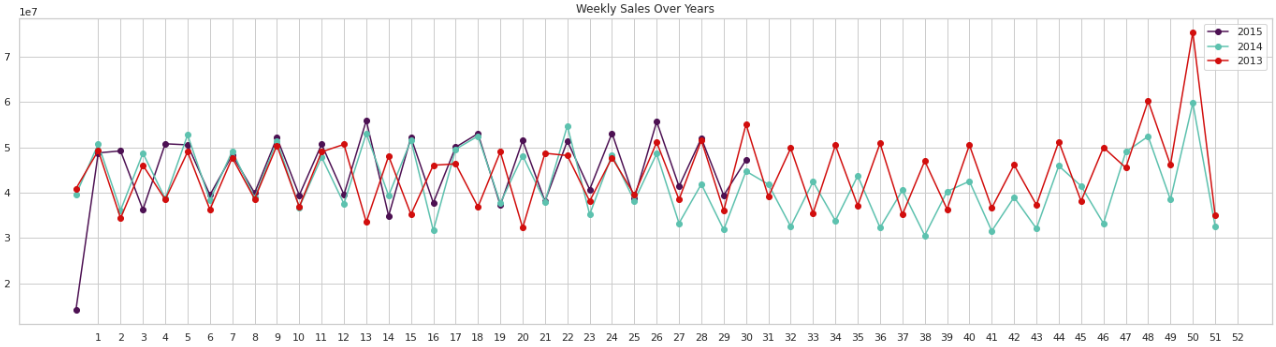

Number of Sales for each Assortment per Store Type


Total Customers

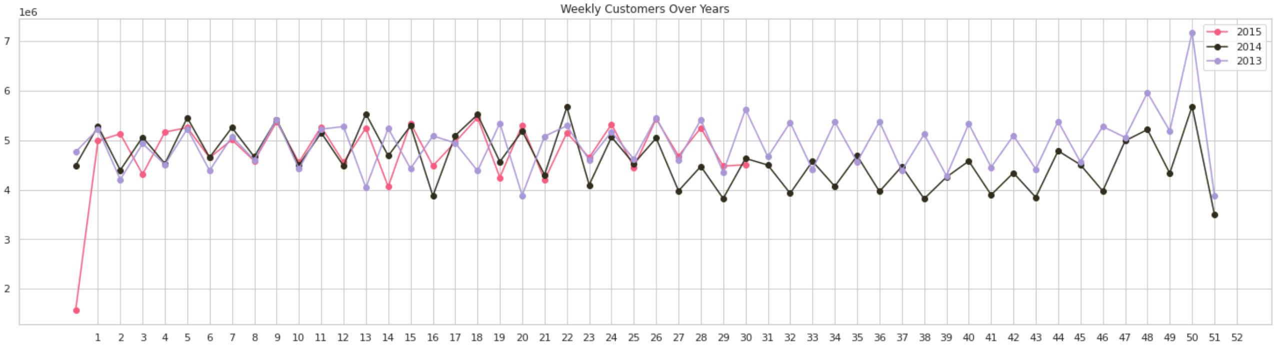

Total Customers per Assortment

Assortment b has the highest average sales and customers. This means assortment b is in high demand. Only store type b sells assortment b.

Exploratory Data Analysis (Contd.)



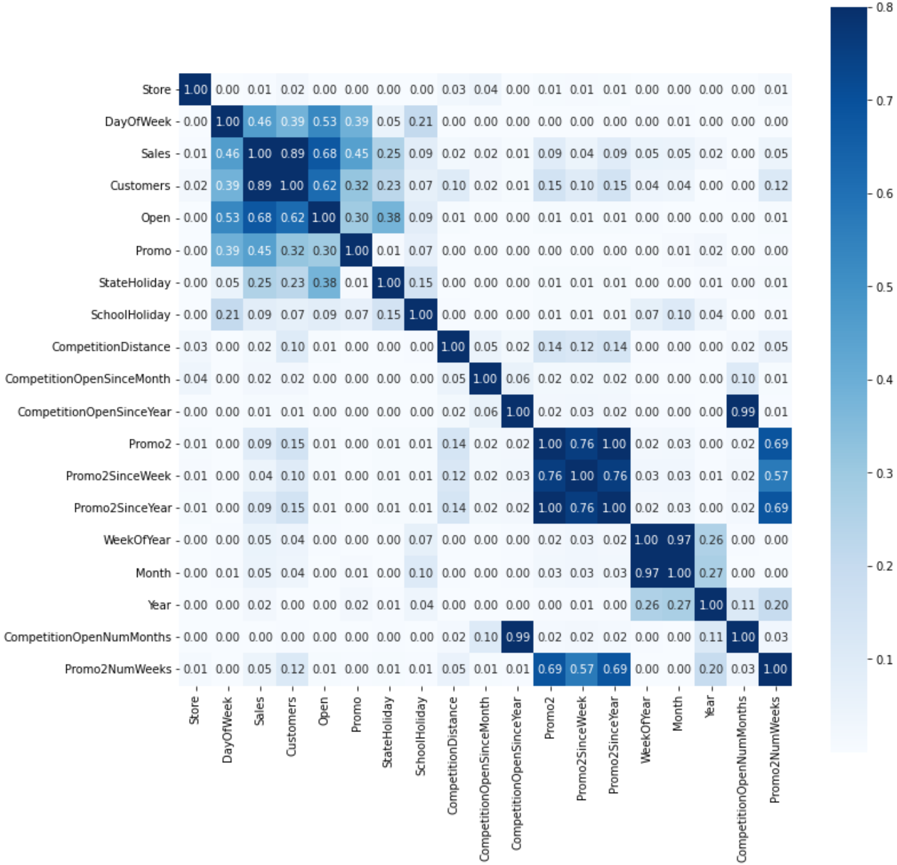
Weekly Sales Over Years



Weekly Customers Over Years

Weekly Sales and Customers are showing almost similar trends. Both peak at mid-December.

Exploratory Data Analysis (Contd.)



	Store	DayOfWeek	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	Promo2	Promo2SinceWeek	Promo2SinceYear	WeekOfYear	Month	Year	CompetitionOpenNumMonths	Promo2NumWeeks
Store	1.00	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.03	0.04	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.01
DayOfWeek	0.00	1.00	0.46	0.39	0.53	0.39	0.05	0.21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
Sales	0.01	0.46	1.00	0.89	0.68	0.45	0.25	0.09	0.02	0.02	0.01	0.09	0.04	0.09	0.05	0.05	0.02	0.00	0.05
Customers	0.02	0.39	0.89	1.00	0.62	0.32	0.23	0.07	0.10	0.02	0.01	0.15	0.10	0.15	0.04	0.04	0.00	0.00	0.12
Open	0.00	0.53	0.68	0.62	1.00	0.30	0.38	0.09	0.01	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.01
Promo	0.00	0.39	0.45	0.32	0.30	1.00	0.01	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00
StateHoliday	0.00	0.05	0.25	0.23	0.38	0.01	1.00	0.15	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.01	0.00	0.01
SchoolHoliday	0.00	0.21	0.09	0.07	0.09	0.07	0.15	1.00	0.00	0.00	0.00	0.01	0.01	0.01	0.07	0.10	0.04	0.00	0.01
CompetitionDistance	0.03	0.00	0.02	0.10	0.01	0.00	0.00	0.00	1.00	0.05	0.02	0.14	0.12	0.14	0.00	0.00	0.00	0.02	0.05
CompetitionOpenSinceMonth	0.04	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.05	1.00	0.06	0.02	0.02	0.02	0.00	0.00	0.00	0.10	0.01
CompetitionOpenSinceYear	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.02	0.06	1.00	0.02	0.03	0.02	0.00	0.00	0.00	0.99	0.01
Promo2	0.01	0.00	0.09	0.15	0.01	0.00	0.01	0.01	0.14	0.02	0.02	1.00	0.76	1.00	0.02	0.03	0.00	0.02	0.69
Promo2SinceWeek	0.01	0.00	0.04	0.10	0.01	0.00	0.01	0.01	0.12	0.02	0.03	0.76	1.00	0.76	0.03	0.03	0.01	0.02	0.57
Promo2SinceYear	0.01	0.00	0.09	0.15	0.01	0.00	0.01	0.01	0.14	0.02	0.02	1.00	0.76	1.00	0.02	0.03	0.00	0.02	0.69
WeekOfYear	0.00	0.00	0.05	0.04	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.02	0.03	0.02	1.00	0.97	0.26	0.00	0.00
Month	0.00	0.01	0.05	0.04	0.00	0.01	0.00	0.10	0.00	0.00	0.00	0.03	0.03	0.03	0.97	1.00	0.27	0.00	0.00
Year	0.00	0.00	0.02	0.00	0.00	0.02	0.01	0.04	0.00	0.00	0.00	0.00	0.01	0.00	0.26	0.27	1.00	0.11	0.20
CompetitionOpenNumMonths	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.10	0.99	0.02	0.02	0.02	0.00	0.00	0.11	1.00	0.03
Promo2NumWeeks	0.01	0.00	0.05	0.12	0.01	0.00	0.01	0.01	0.05	0.01	0.01	0.69	0.57	0.69	0.00	0.00	0.20	0.03	1.00

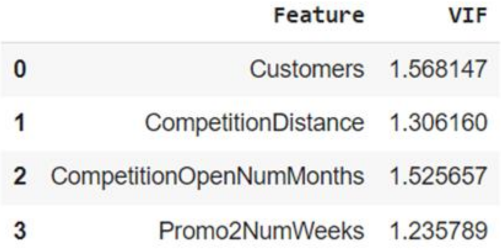
- ‘Customers’ and ‘Sales’ are highly correlated.
- We can see that ‘WeekOfYear’ and ‘Month’ are also highly correlated.
- ‘Open’ is moderately correlated with ‘Sales’ and ‘Customers’.
- ‘Promo2’ is correlated with ‘Promo2SinceWeek’ and ‘Promo2SinceYear’.
- ‘CompetitionOpenNumMonths’ is highly correlated with ‘CompetitionOpenSinceYear’.
- ‘Promo2NumWeeks’ is moderately correlated with ‘Promo2’, ‘Promo2SinceWeek’ and ‘Promo2SinceYear’.

Feature Engineering

- ‘Store’ is dropped since sales can be predicted through store type, assortment, etc.
- ‘Date’ is also dropped since there is already day of week and week of year features in the dataset.
- ‘CompetitionOpenSinceMonth’ and ‘CompetitionOpenSinceYear’ are dropped as the information provided by them can be obtained from ‘CompetitionOpenNumMonths’.
- ‘Promo2’, ‘Promo2SinceWeek’ and ‘Promo2SinceYear’ are also dropped as the information provided by them can be obtained from ‘Promo2NumWeeks’.
- ‘Month’ is dropped since we get the same information from ‘WeekOfYear’.
- ‘Year’ is also dropped as we have already established in EDA that it’s not the year that influence the sales but ‘DayOfWeek’ and ‘WeekOfYear’.

Feature Engineering (Contd.)

- The variance inflation factor (VIF) of all numerical features except 'Sales' is calculated in order to remove highly correlated features.
- Features having VIF greater than 5 should be eliminated.

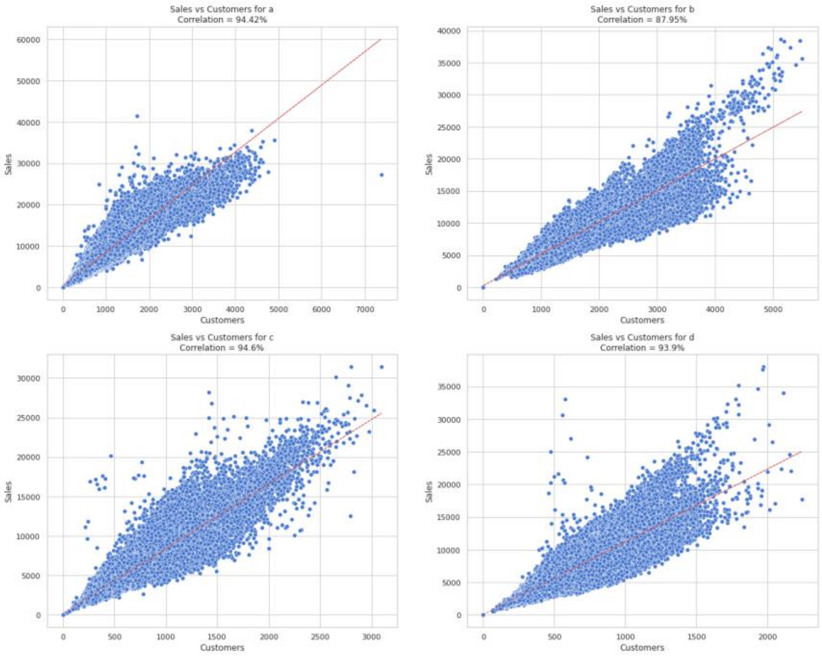


	Feature	VIF
0	Customers	1.568147
1	CompetitionDistance	1.306160
2	CompetitionOpenNumMonths	1.525657
3	Promo2NumWeeks	1.235789

- All features have VIF less than 5.

Feature Engineering (Contd.)

Before removing outliers in ‘StoreType’

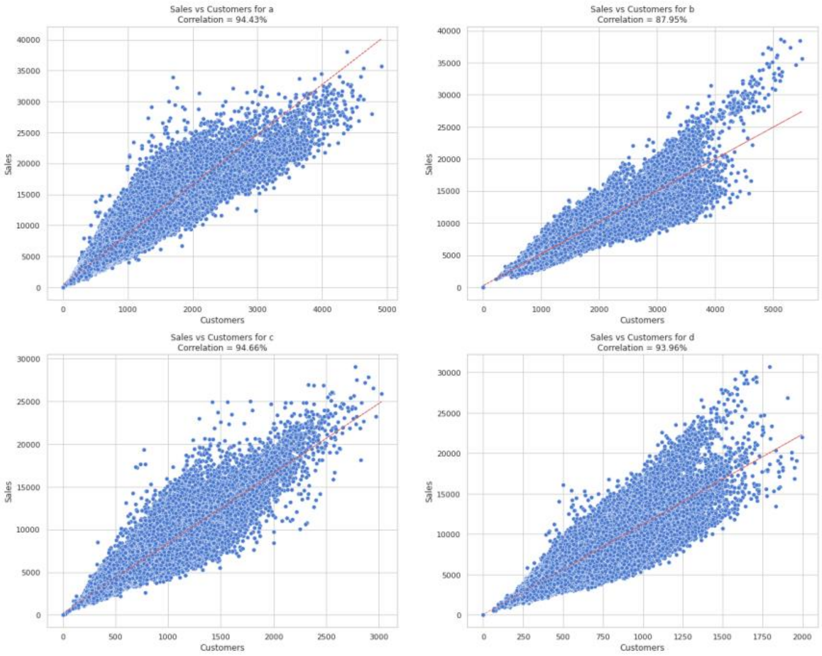

Sales vs Customers for a
Correlation = 94.42%

Sales vs Customers for b
Correlation = 87.95%

Sales vs Customers for c
Correlation = 94.6%

Sales vs Customers for d
Correlation = 93.9%

After removing outliers in ‘StoreType’


Sales vs Customers for a
Correlation = 94.43%

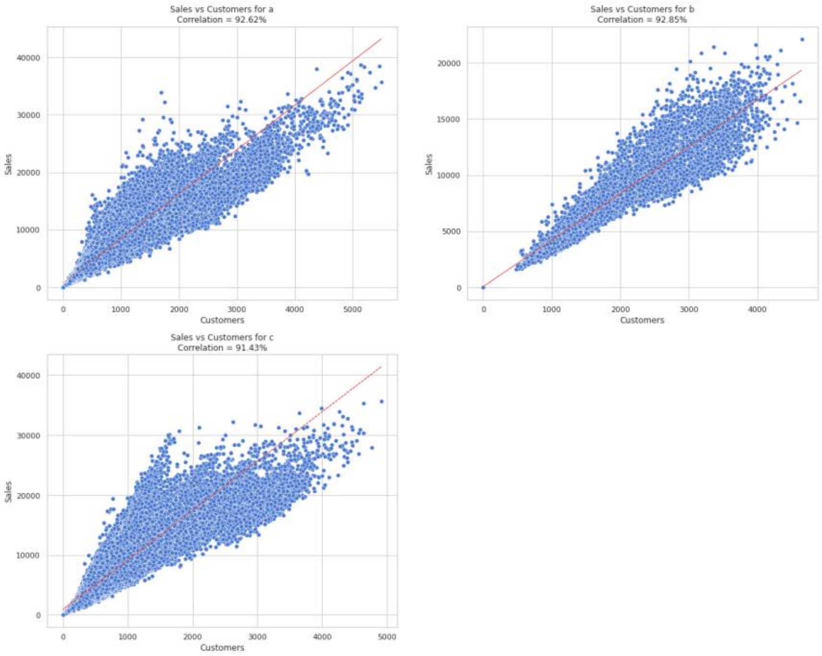
Sales vs Customers for b
Correlation = 87.95%

Sales vs Customers for c
Correlation = 94.66%

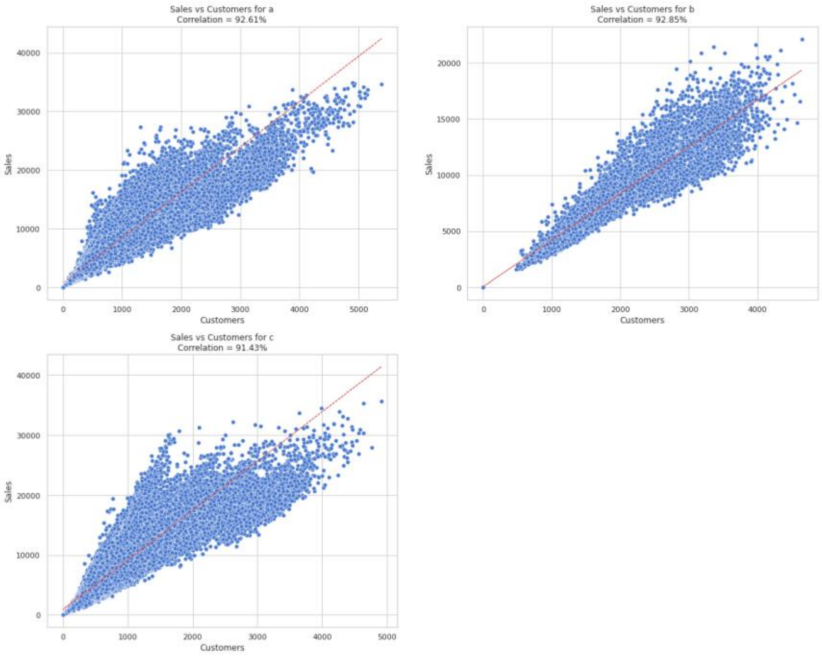
Sales vs Customers for d
Correlation = 93.96%

Feature Engineering (Contd.)

Before removing outliers in ‘Assortment’

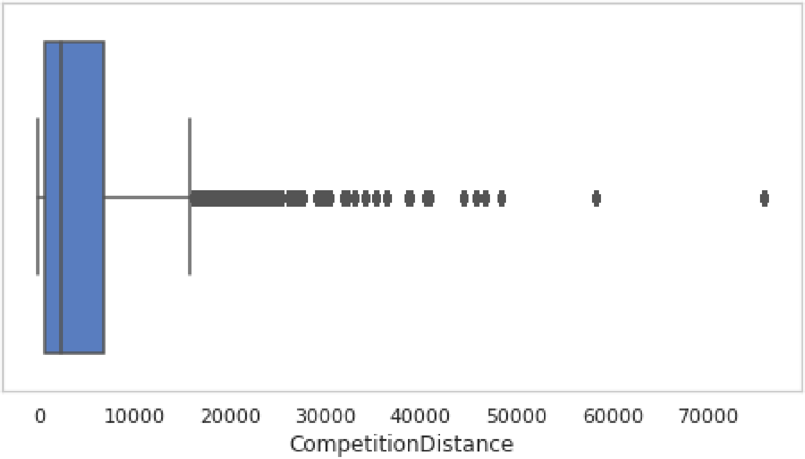


After removing outliers in ‘Assortment’

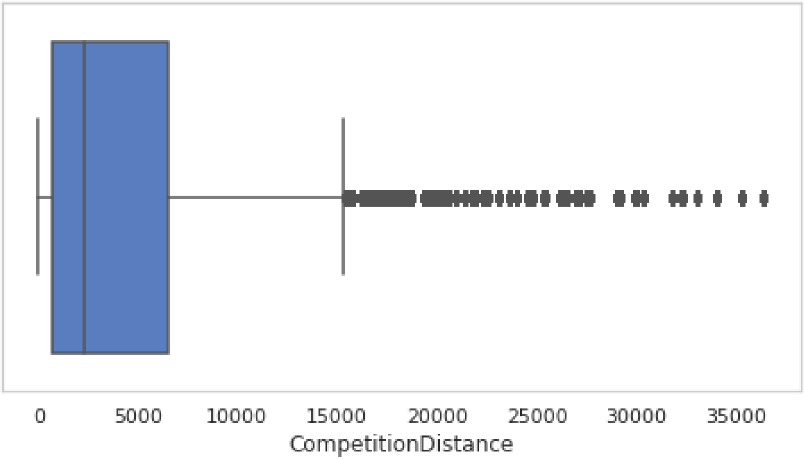


Feature Engineering (Contd.)

Before removing outliers in ‘CompetitionDistance’

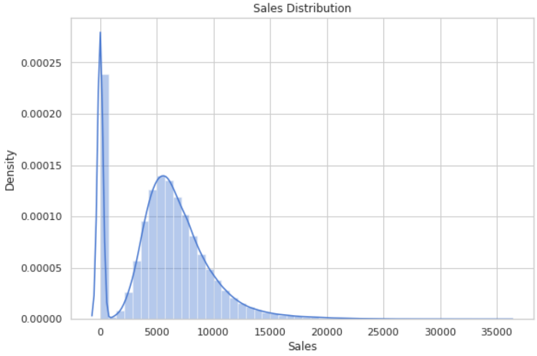


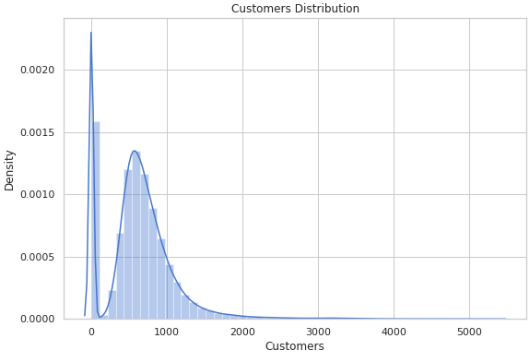
After removing outliers in ‘CompetitionDistance’

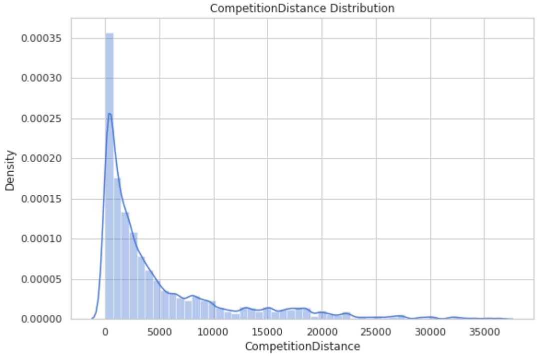


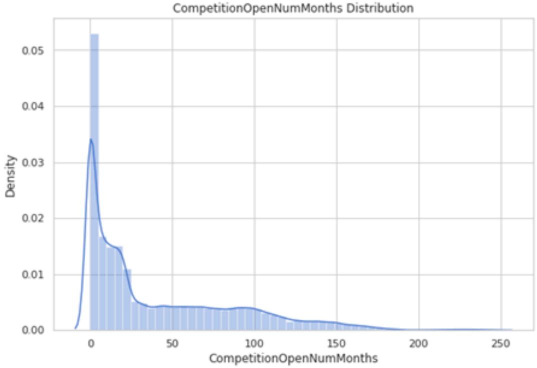
- Almost 10% of data are outliers and removing them will cause the loss of useful information.
- So, only the outliers having values above 37500 were removed.

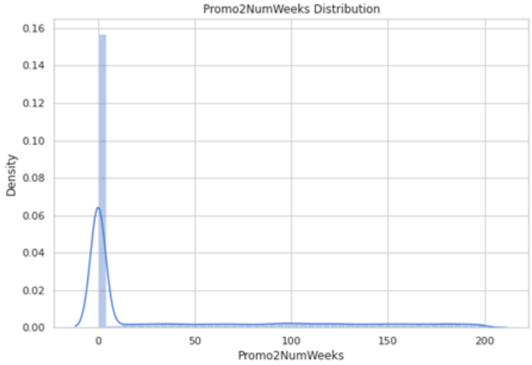
Feature Engineering (Contd.)


Sales Distribution


Customers Distribution

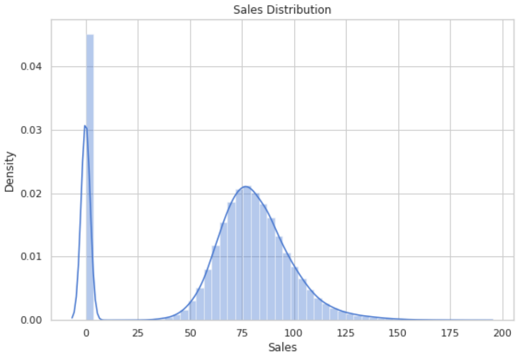

CompetitionDistance Distribution

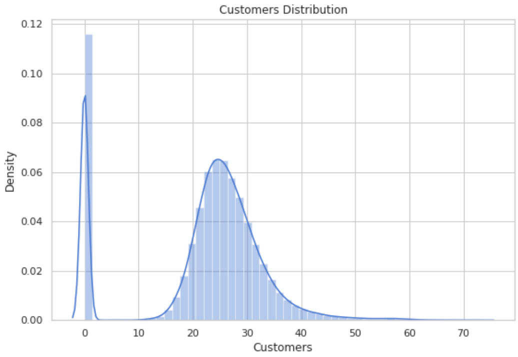

CompetitionOpenNumMonths Distribution

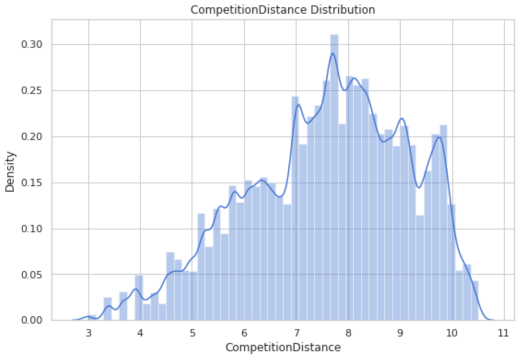

Promo2NumWeeks Distribution

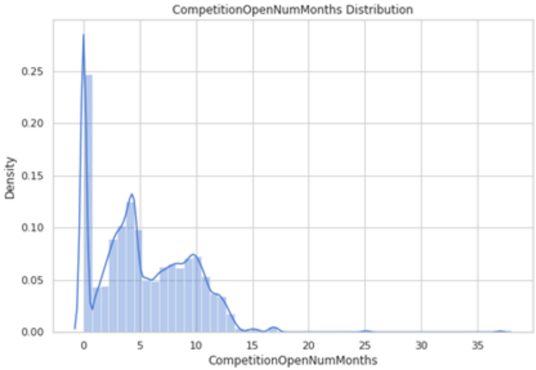
All features are positively skewed but linear regression assumes normal distribution.

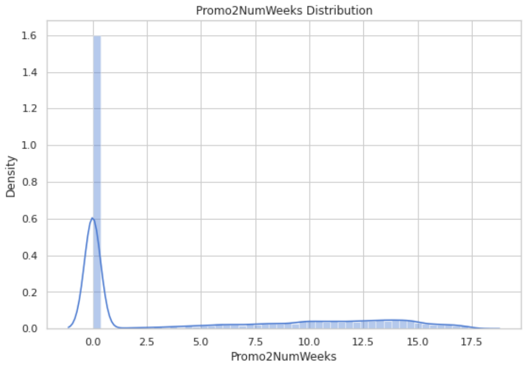
Feature Engineering (Contd.)


Sales Distribution


Customers Distribution


CompetitionDistance Distribution


CompetitionOpenNumMonths Distribution


Promo2NumWeeks Distribution

‘Sales’, ‘Customers’, ‘CompetitionOpenNumMonths’ and ‘Promo2NumWeeks’ are square root transformed and ‘CompetitionDistance’ is log transformed.

Feature Engineering (Contd.)

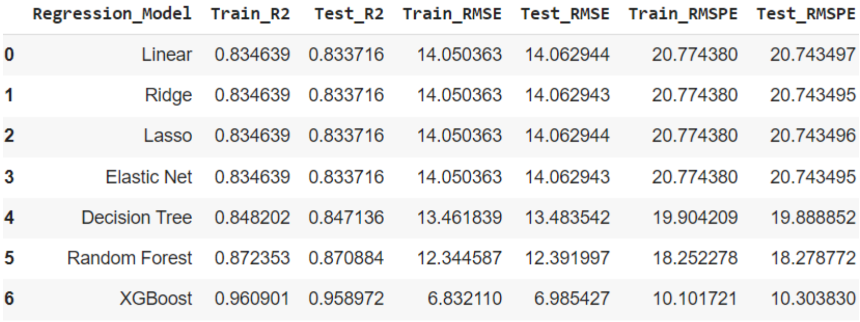
- Most algorithms cannot handle the categorical variables unless they are converted into a numerical value.
- ‘StoreType’ and ‘Assortment’ were encoded using one hot encoder while ‘PromoInterval’ was dummified.
- To overcome dummy variable trap, one resultant feature from each encoded feature must be removed. Correlation matrix was used to decide which features to remove.
- ‘PromoInterval_Jan,Apr,Jul,Oct’, ‘StoreType_c’ & ‘Assortment_b’ have the least correlation to Sales and they were removed.

Modelling

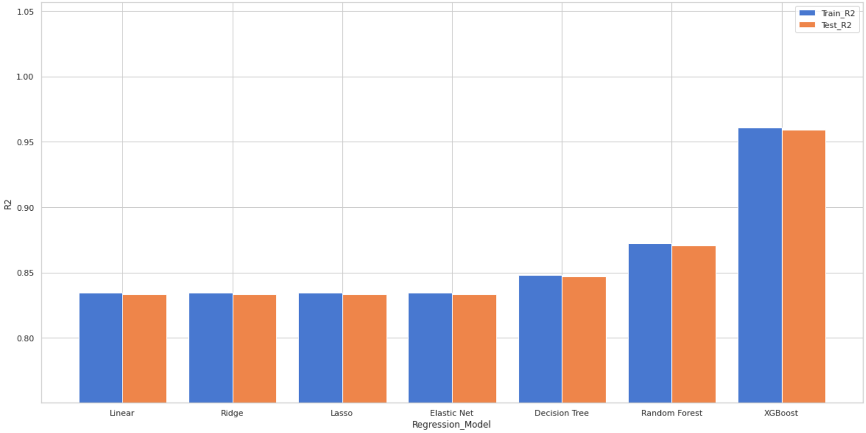
- Input and target data were separated, and both were split into training and test data with 25% test data.
- The feature ‘Customers’ was also removed since the number of customers for the period, which is under consideration for forecasting, won't be available until the mentioned period is over.
- Training and test data of independent features were scaled using standardization.
- Model training was done with these data using 7 different algorithms:

  1. Linear regression
  2. Ridge regression
  3. Lasso regression
  4. Elastic net regression
  5. Decision tree regression
  6. Random forest regression
  7. XGBoost regression

Modelling (Contd.)



	Regression_Model	Train_R2	Test_R2	Train_RMSE	Test_RMSE	Train_RMSPE	Test_RMSPE
0	Linear	0.834639	0.833716	14.050363	14.062944	20.774380	20.743497
1	Ridge	0.834639	0.833716	14.050363	14.062943	20.774380	20.743495
2	Lasso	0.834639	0.833716	14.050363	14.062944	20.774380	20.743496
3	Elastic Net	0.834639	0.833716	14.050363	14.062943	20.774380	20.743495
4	Decision Tree	0.848202	0.847136	13.461839	13.483542	19.904209	19.888852
5	Random Forest	0.872353	0.870884	12.344587	12.391997	18.252278	18.278772
6	XGBoost	0.960901	0.958972	6.832110	6.985427	10.101721	10.303830



- Evaluation metrics like R-squared, RMSE, RMSPE, etc. were calculated for each model.
- R2 score can be used to compare different models and find out which one gives higher accuracy. Higher the R2 score, higher the accuracy.
- The model built using XGBoost algorithm has the highest R2 score with ~96%, followed by the one using random forest and decision tree.

Conclusion

EDA Conclusions

- Mondays have most sales since most of the Sundays are closed.
- Promotions seem to have a significant effect on sales but not for the number of customers.
- Store type b has higher sales and customers per store than other store types.
- Assortment b is available only at store type b and it has more average sales and customers than any other assortment.
- Weekly sales and customers peak at the mid-December. Best guess is that people buy drugs in advance just before the shops close for the holidays.

Modelling Conclusions

- The model built using XGBoost algorithm gives unusually high accuracy. This may lead to overfitting. Therefore, it is advisable to not use this model.
- Among the remaining, the model built using random forest algorithm is the most accurate one.
- If model interpretability is more important than accuracy, model built using decision tree algorithm should be chosen over the one using random forest algorithm.
- Decision tree based algorithms are slightly more accurate than linear regression based algorithms.