Sales Prediction Report

C3t3 Regression model –
Sugitha Devarajan

Problem

- To analyze historical sales data and then make sales volume predictions for a list of new product types. This will help the sales team better understand how types of products might impact sales across the enterprise.
- Predicting sales of four different product types: PC, Laptops, Netbooks and Smartphones
- Assessing the impact services reviews and customer reviews have on sales of different product types.

Data

- Data is a small set with following features
  - Product Type
  - Product number
  - Price
  - Product Reviews
  - Customer service review
  - Recommended product
  - Best seller Rank
  - Shipping Weight
  - Product Depth , depth, and height
  - Profit Margin
  - Volume – Dependent variable

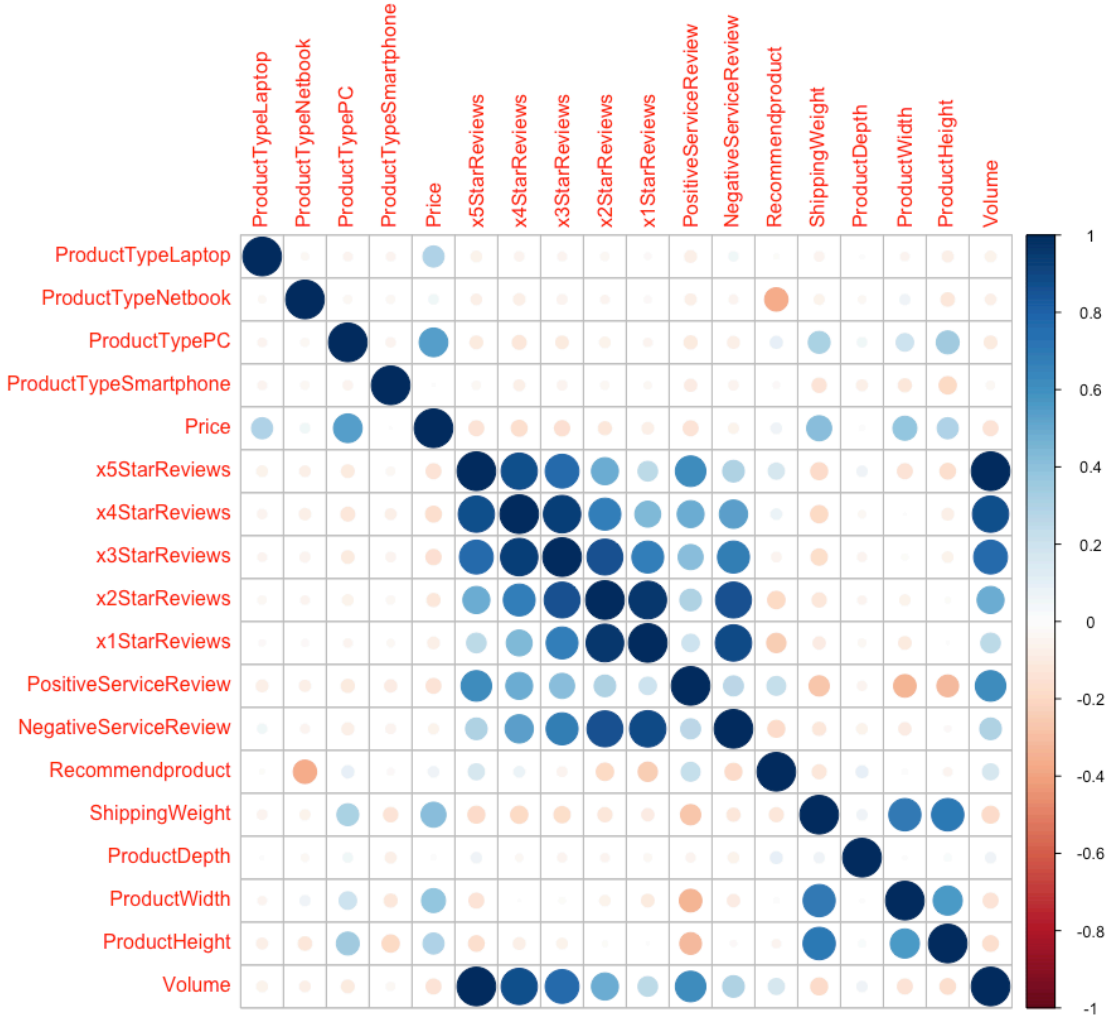
```
> str(df_ps)
'data.frame':   80 obs. of  18 variables:
 $ ProductType          : chr  "PC" "PC" "PC" "Laptop" ...
 $ ProductNum           : int  101 102 103 104 105 106 107 108 109 110 ...
 $ Price                : num  949 2250 399 410 1080 ...
 $ x5StarReviews        : int  3 2 3 49 58 83 11 33 16 10 ...
 $ x4StarReviews        : int  3 1 0 19 31 30 3 19 9 1 ...
 $ x3StarReviews        : int  2 0 0 8 11 10 0 12 2 1 ...
 $ x2StarReviews        : int  0 0 0 3 7 9 0 5 0 0 ...
 $ x1StarReviews        : int  0 0 0 9 36 40 1 9 2 0 ...
 $ PositiveServiceReview: int  2 1 1 7 7 12 3 5 2 2 ...
 $ NegativeServiceReview: int  0 0 0 8 20 5 0 3 1 0 ...
 $ Recommendproduct     : num  0.9 0.9 0.9 0.8 0.7 0.3 0.9 0.7 0.8 0.9 ...
 $ BestSellersRank      : int  1967 4806 12076 109 268 64 NA 2 NA 18 ...
 $ ShippingWeight       : num  25.8 50 17.4 5.7 7 1.6 7.3 12 1.8 0.75 ...
 $ ProductDepth         : num  23.9 35 10.5 15 12.9 ...
 $ ProductWidth         : num  6.62 31.75 8.3 9.9 0.3 ...
 $ ProductHeight        : num  16.9 19 10.2 1.3 8.9 ...
 $ ProfitMargin         : num  0.15 0.25 0.08 0.08 0.09 0.05 0.05 0.05 0.05 0.05 ...
 $ Volume               : int  12 8 12 196 232 332 44 132 64 40 ...
>
```

Data Pre-Processing

- Checked for “na” – Best Sellers Rank had lot of “na” so was dropped.
- No duplicates
- Product number was dropped
- Product type a categorical column was converted using dummy variables to numerical
- Profit Margin and other unnecessary product type was dropped after checking correlation.
- Volume is the dependent variable

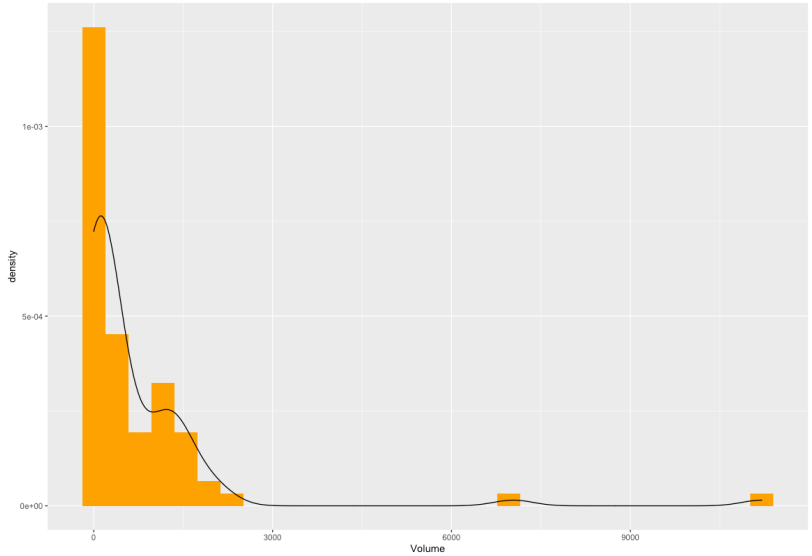
Correlation Map for selected features

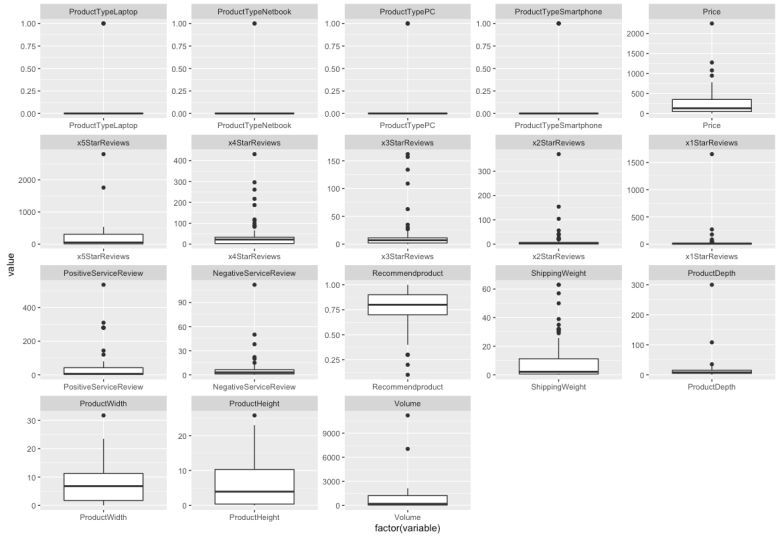
- Seems the customer product review have high correlation with Volume



Exploratory Data Analysis

- The Volume is not distributed evenly
- box plot of all selected features for finding outliers – there were some in almost all features





Approach to Model

- The data was split in to 80 and 20 for training and testing.
- Linear model, SVM , Random Forest, and Gradient Boosting regression models are selected to predict the Volume.
- In regression models, the main aim is to make sure we fit a model with high R squared, low Root Mean Squared Error, and low Mean Absolute Error

Linear Regression Model

- A linear regression is a statistical model that analyzes the relationship between a response variable (often called y) and one or more variables and their interactions (often called x or explanatory variables).
- Linear model did not perform well in sales volume prediction – gave a 100% R-Squared meaning it was overfitting or unreliable

	R2	RMSE	MAE
1	1	6.529146e-13	4.8305e-13

- The reason for perfectly fitted r squared is the model was not right for non-parametric data type. So, moving on to next models.

Support Vector Machine SVM

- The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.
- SVM regression is considered a nonparametric technique because it relies on kernel functions.
- Using EPS regression and NU regression type along with different kernel i.e., Linear, Polynomial, Sigmoid, and radial I was able to built 8 models with default control.
- SVM linear NU regression model gave 0.99 in R Squared for test and train data.

	method	R2	RMSE	MAE
1	svm Linear NU	0.9999998	1.076049	0.6441106
2	svm Ploy NU	0.8495258	1303.338709	570.4159118
3	svm radial NU	0.5309261	1409.531715	507.7145664
4	svm sigmoid NU	0.2379326	1547.905056	616.8033844
5	svm Linear EPS	0.9900156	194.399565	123.7953857
6	svm Ploy EPS	0.8621021	1294.806469	571.9549618
7	svm radial EPS	0.5464925	1418.886163	536.0353591
8	svm sigmoid EPS	0.2755657	1513.418822	626.4203229

```
> summary(ps_svmlinearnu)

Call:
svm(formula = Volume ~ ., data = ps_training, type = "nu-regression", kernel = "linear", cross = 3)


Parameters:
   SVM-Type:  nu-regression
 SVM-Kernel:  linear
       cost:  1
         nu:  0.5

Number of Support Vectors:  47



3-fold cross-validation on training data:

Total Mean Squared Error: 92.94487
Squared Correlation Coefficient: 0.9999568
Mean Squared Errors:
 0.6528951 45.6804 228.3062



> dfsl<-data.frame( method="svm Linear NU",R2 = R2(predictlinearnu_ps, ps_testing$Volume),
+                   RMSE = RMSE(predictlinearnu_ps, ps_testing$Volume),
+                   MAE = MAE(predictlinearnu_ps, ps_testing$Volume))
> dfsl
          method       R2     RMSE       MAE
1 svm Linear NU 0.9999998 1.076049 0.6441106
```

Random Forest

- The random forest algorithm works by aggregating the predictions made by multiple decision trees of varying depth. Every decision tree in the forest is trained on a subset of the dataset called the bootstrapped dataset.
- I used two methods of random forest. RF and PARRF.
- Rf

	R2	RMSE	MAE
1	0.8645149	1052.192	329.7855

- PARRF

	R2	RMSE	MAE
1	0.7293347	1366.312	572.9089

- RF Method seems to have higher R-squared.

Gradient Boost

- Gradient boosted machines (GBMs) are an extremely popular machine learning algorithm that have proven successful across many domains and is one of the leading methods for winning Kaggle competitions. Whereas random forests build an ensemble of deep independent trees, GBMs build an ensemble of shallow and weak successive trees with each tree learning and improving on the previous. When combined, these many weak successive trees produce a powerful “committee” that are often hard to beat with other algorithms
- Predicted data

	R2	RMSE	MAE
1	0.3718745	1382.971	599.6776

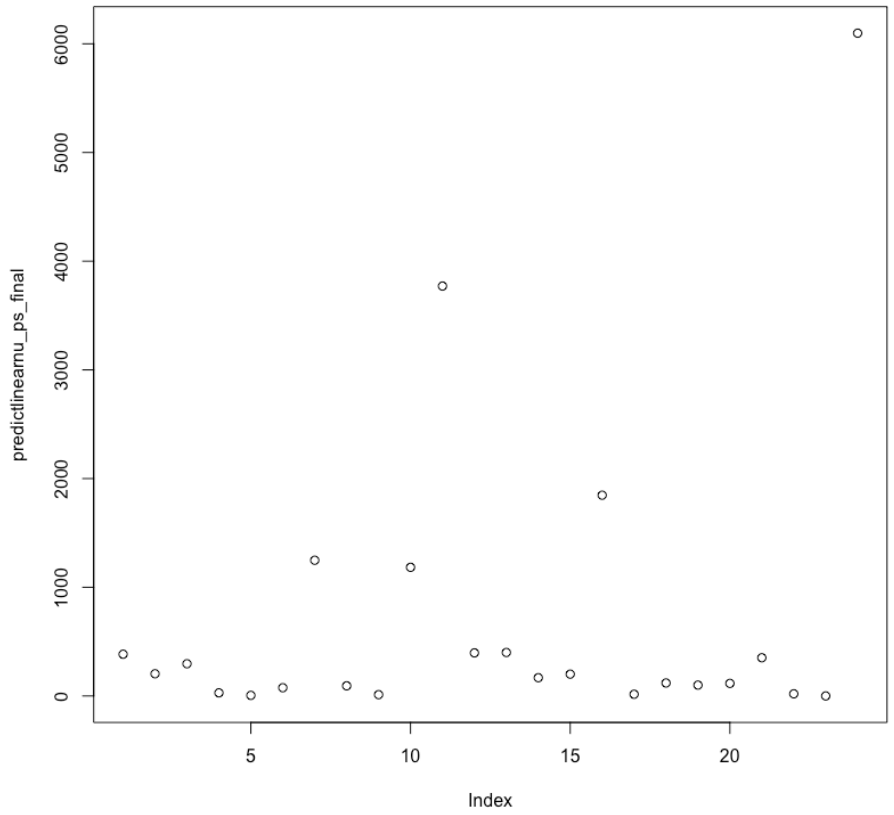
Select Model for Prediction

- With 11 different models build the closest which fit the best with high R2, low RMSE , and low MAE is **SVM Linear NU regression model**.

	method	R2	RMSE	MAE
1	svm Linear NU	0.9999998	1.076049	0.6441106
2	svm Ploy NU	0.8495258	1303.338709	570.4159118
3	svm radial NU	0.5309261	1409.531715	507.7145664
4	svm sigmoid NU	0.2379326	1547.905056	616.8033844
5	svm Linear EPS	0.9900156	194.399565	123.7953857
6	svm Ploy EPS	0.8621021	1294.806469	571.9549618
7	svm radial EPS	0.5464925	1418.886163	536.0353591
8	svm sigmoid EPS	0.2755657	1513.418822	626.4203229
9	rf	0.8645149	1052.192074	329.7854756
10	parf	0.7293347	1366.312052	572.9089385
11	gbm	0.3718745	1382.971138	599.6776426

Predicting Sales for new product csv

- With SVM Linear NU regression model the sales volume was predicted and plotted.
- The output was extracted in a csv file.



Conclusion

- **Did you learn anything of potential business value from this analysis?**

During the modeling process I learnt that customer reviews with 5 star is highly correlated with volume along with positive service reviews.

- **Was it straightforward to rerun your projections of sales volume using both models?**

No, the data was processed exactly like the training set to predict the sales volume using the models built.

- **What are the main lessons you've learned from this experience?**

Data distribution is very important in model building especially when it comes to regression models.

- **What recommendations would you give to the sales department regarding your findings relating to the different types of reviews?**

Reviews are very important for sales volume increase more than price or product size. Both Customer Product review and Service review needs to be highly focused for best sales volume.