Customer Brand Preferences - Course 3, Task 3

Blackwell Electronics Investigation

Primary Objective:

- Predict sales of four different product types: PC, Laptops, Netbooks and Smartphones
- Asses the impact services reviews and customer reviews have on sales of different product types

Business Question:

Can we accurately predict sales volumes of different product types to better understand how new products might impact future sales?

Data Set

For this analysis and modeling we used a existing product data set with the sales volumes of various product types and product IDs, combined with collected attributes for each. We are also provided a data set of new products with the same attributes, though without volume data. We'll build a model using the existing product data set and apply to the new product data set to predict future sales volumes for Blackwell Electronics.

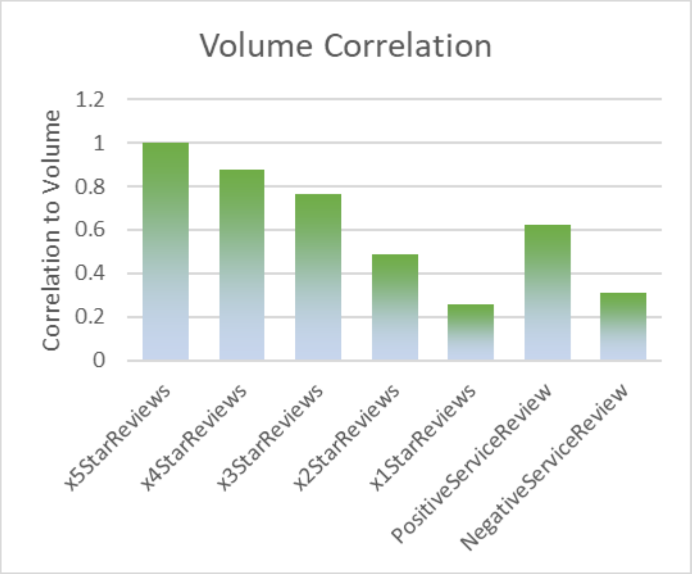


Figure 1: Correlation values between the customer and service reviews to the sales volume for existing products

Feature Engineering

Correlation Analysis

In the correlation analysis we're able to asses the impact of services reviews and customer reviews. As shown in the truncated correlation plot of Figure 1, the more positive star and service reviews have stronger correlations to the sales volume. However, even 1-star reviews and negative service reviews have positive correlations – indicating that higher sales volumes simply result in more reviews, good or bad.

A large issue with this data set is that 5-star reviews correlate to the sales volume at 100%. It is discussed further in the recommendation section, but per the guidance of correlation analysis, we should remove this independent variable from our data set. Second, two of the independent variables – 3-star and 2-star reviews are highly correlated (0.93), so the 2-star reviews variable is removed. This new tuned data set is saved and is used for model generation.

Recursive Feature Elimination (RFE)

Once the 5-star Reviews (100% correlation) was removed, RFE was run for a random forest model. With this, only two variables were deemed necessary to make good predictions – PositiveServiceReview and x4StarReviews.

Feature Construction

To better capture the customer reviews, we created a Star score that averaged the score for each product, resulting in a value between 1 and 5 (much like any other online product star rating) to better capture product satisfaction. Similarly, we created a Service Score (0-1) to estimate the quality of the service for each product (Pos Score / (Pos Score + Neg Score)).

Feature Removal

The BestSellersRank attribute was missing data for several rows. Rather than ignore/delete entire product IDs, it made more sense to eliminate this incomplete variable from our data set, given the large number of other variables and relatively limited row count in the source data.

Predictive Modelling

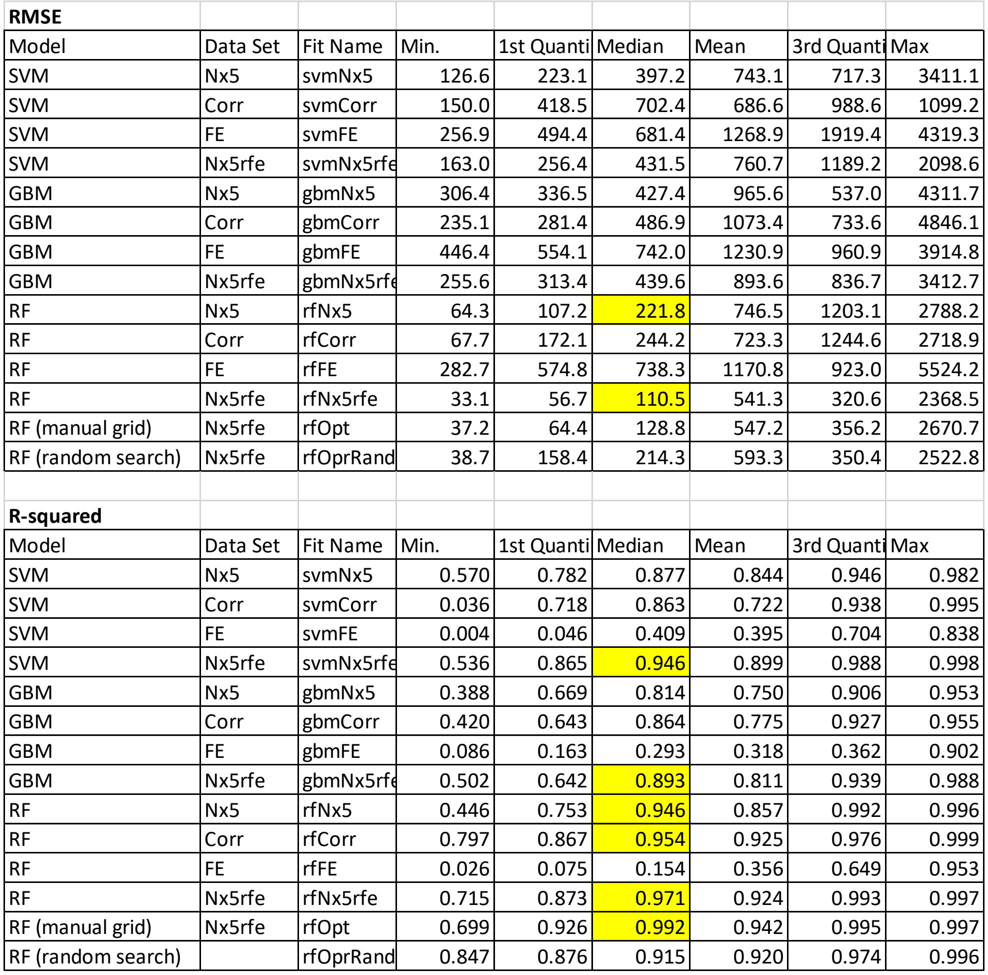
Not surprisingly, many models return warnings and errors when the dataframe includes the 100% correlated 5-star review variable. As such, this variable is removed from the main data set (dfNx5 = df with No 5-star). If it is indeed realistic, there is no need for any model generation.

To fully capture the optimum fit with this data set, we performed a cross product of 3 main models (SVM, GBM, RF) with the 4 generated data sets

- **Nx5** - Out-of-box with 5-star removed
- **Corr** – 5-star and 2-star removed
- **FE** – feature construction: average star review score and average service review score with all other star-score and service-score variables removed
- **Nx5rfe** – 5-star removed and limited to the recursive feature engineering results: keeping only 4-star reviews and positive service reviews
  with 3 main models (SVM, GBM, RF)

Following this analysis – the leading RF model was tuned using manual grid and random search to produce to more fit candidates. The compiled Training data fits are shown in Table 1.

Table 1: Training fits for various models and factors



RMSE

Model	Data Set	Fit Name	Min.	1st Quanti	Median	Mean	3rd Quanti	Max
SVM	Nx5	svmNx5	126.6	223.1	397.2	743.1	717.3	3411.1
SVM	Corr	svmCorr	150.0	418.5	702.4	686.6	988.6	1099.2
SVM	FE	svmFE	256.9	494.4	681.4	1268.9	1919.4	4319.3
SVM	Nx5rfe	svmNx5rfe	163.0	256.4	431.5	760.7	1189.2	2098.6
GBM	Nx5	gbmNx5	306.4	336.5	427.4	965.6	537.0	4311.7
GBM	Corr	gbmCorr	235.1	281.4	486.9	1073.4	733.6	4846.1
GBM	FE	gbmFE	446.4	554.1	742.0	1230.9	960.9	3914.8
GBM	Nx5rfe	gbmNx5rfe	255.6	313.4	439.6	893.6	836.7	3412.7
RF	Nx5	rfNx5	64.3	107.2	221.8	746.5	1203.1	2788.2
RF	Corr	rfCorr	67.7	172.1	244.2	723.3	1244.6	2718.9
RF	FE	rfFE	282.7	574.8	738.3	1170.8	923.0	5524.2
RF	Nx5rfe	rfNx5rfe	33.1	56.7	110.5	541.3	320.6	2368.5
RF (manual grid)	Nx5rfe	rfOpt	37.2	64.4	128.8	547.2	356.2	2670.7
RF (random search)	Nx5rfe	rfOprRand	38.7	158.4	214.3	593.3	350.4	2522.8

R-squared

Model	Data Set	Fit Name	Min.	1st Quanti	Median	Mean	3rd Quanti	Max
SVM	Nx5	svmNx5	0.570	0.782	0.877	0.844	0.946	0.982
SVM	Corr	svmCorr	0.036	0.718	0.863	0.722	0.938	0.995
SVM	FE	svmFE	0.004	0.046	0.409	0.395	0.704	0.838
SVM	Nx5rfe	svmNx5rfe	0.536	0.865	0.946	0.899	0.988	0.998
GBM	Nx5	gbmNx5	0.388	0.669	0.814	0.750	0.906	0.953
GBM	Corr	gbmCorr	0.420	0.643	0.864	0.775	0.927	0.955
GBM	FE	gbmFE	0.086	0.163	0.293	0.318	0.362	0.902
GBM	Nx5rfe	gbmNx5rfe	0.502	0.642	0.893	0.811	0.939	0.988
RF	Nx5	rfNx5	0.446	0.753	0.946	0.857	0.992	0.996
RF	Corr	rfCorr	0.797	0.867	0.954	0.925	0.976	0.999
RF	FE	rfFE	0.026	0.075	0.154	0.356	0.649	0.953
RF	Nx5rfe	rfNx5rfe	0.715	0.873	0.971	0.924	0.993	0.997
RF (manual grid)	Nx5rfe	rfOpt	0.699	0.926	0.992	0.942	0.995	0.997
RF (random search)		rfOprRand	0.847	0.876	0.915	0.920	0.974	0.996

Prediction

Upon prediction of the highest performing models on the arbitrarily selected Test data segment, we find much lower R-squared scores and higher RMSE values. This may partially be due to overfitting or the small number of samples in the training data set. Also troublesome was that some of the models produced negative volume numbers, invalidating those models. Upon re-reviewing the models based on the predicted values of the Test data set, the GBM model with the Nx5rfe data set proved most useful, with an RMSE of 592 and an R-squared of 0.808. Comparing the Ground Truth values to these models illustrates that the RF models tend to dramatically overestimate the volume sales of some of the higher volume items as shown by the dots in the top right of Figure 2.

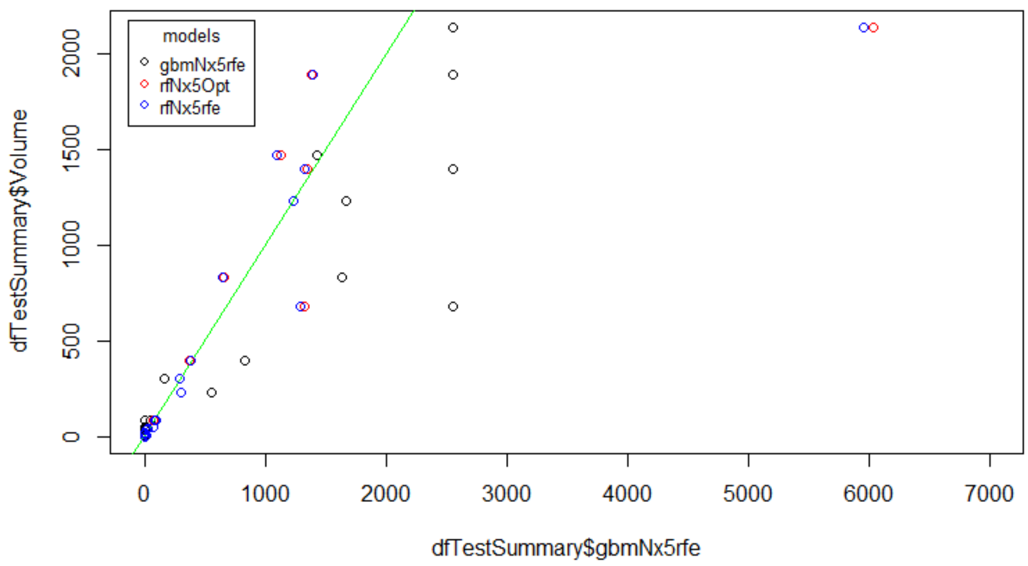


Figure 2: Predicted versus actual values for various models. The green line represents a perfectly accurate prediction (prediction equals ground truth)

Using this recursive feature reduction and the GBM model on the New Product data set, we're able to make the predictions seen in Table 2. Unfortunately, now the GBM model has produced a negative volume prediction for one of the smartphone products, so for comparison, the results from the RF model using the same RFE data set are also included, and look promising.

Table 2: Predicted volumes for new products

ProductType	ProductNum	Price	x5StarReviews	x4StarReviews	PositiveServiceReview	Volume Prediction GBM	Volume Prediction RF	4X 5-star
PC	171	699	96	26	12	111	390	384
PC	172	860	51	11	7	51	144	204
Laptop	173	1199	74	10	11	337	188	296
Laptop	175	1199	7	2	2	10	53	28
Laptop	176	1999	1	1	0	10	8	4
Netbook	178	399.99	19	8	2	10	49	76
Netbook	180	329	312	112	28	2250	1211	1248
Netbook	181	439	23	18	5	25	116	92
Netbook	183	330	3	4	1	10	27	12
Smartphor	193	199	99	26	8	-50	296	396
Smartphor	194	49	100	26	14	759	527	400
Smartphor	195	149	42	8	4	10	80	168
Smartphor	196	300	50	19	5	25	143	200

Recommendations

Due to the perfect correlation between 5-star reviews for a given product number and the sales volume, it would be good to scrutinize the data source. It appears unrealistic that product sales would consistently be exactly 4X the number of 5-star reviews. Should this data be determined to be trustworthy, then no advanced modeling is needed, but this feels like an unlikely outcome. We can build models to predict future sales volumes, but without more rigorous testing, it is hard to put too much faith in these models.

Lessons Learned

A challenge with this use of predictive analytics for this task is that the training data set is very small. There was a fairly large discrepancy between training set and test set results, which I suspect may stem from the fairly limited training data. I also wonder if RFE should be done separately for each model being investigated – or conversely, if I should have started with the base models with the out-of-the-box data set and then considered RFE only on the most promising ones. It does appear that the RFE constructed using random forest improved the prediction capability for SVM, GBM, and RF though (Nx5 vs. Nx5rfe for each)