Hardik Shahu

Predicting Sales

Overview of the Tasks

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.

Gathering the Data

For this project, we were given two sets of similar data; one had historical data on existing products and another one was on new products.

The attributes in both were the same, some of them consisted of:

- Product Type
- Price
- Reviews/ Review stars
- Product dimensions
- Profit margins
- Volume sold (only for the historical dataset)

Looking at the Data

This is a small snipped of the data:

ProductTy	ProductNu	Price	x5StarRevi	x4StarRevi	x3StarRevi	x2StarRevi	x1StarRevi	PositiveSer	NegativeSe	Recommer	BestSellers	ShippingW	ProductDe	ProductWi	ProductHe	ProfitMarg	Volume
PC	101	949	3	3	2	0	0	2	0	0.9	1967	25.8	23.94	6.62	16.89	0.15	12
PC	102	2249.99	2	1	0	0	0	1	0	0.9	4806	50	35	31.75	19	0.25	8
PC	103	399	3	0	0	0	0	1	0	0.9	12076	17.4	10.5	8.3	10.2	0.08	12
Laptop	104	409.99	49	19	8	3	9	7	8	0.8	109	5.7	15	9.9	1.3	0.08	196
Laptop	105	1079.99	58	31	11	7	36	7	20	0.7	268	7	12.9	0.3	8.9	0.09	232
Accessorie	106	114.22	83	30	10	9	40	12	5	0.3	64	1.6	5.8	4	1	0.05	332
Accessorie	107	379.99	11	3	0	0	1	3	0	0.9	NA	7.3	6.7	10.3	11.5	0.05	44
Accessorie	108	65.29	33	19	12	5	9	5	3	0.7	2	12	7.9	6.7	2.2	0.05	132
Accessorie	109	119.99	16	9	2	0	2	2	1	0.8	NA	1.8	10.6	9.4	4.7	0.05	64
Accessorie	110	16.99	10	1	1	0	0	2	0	0.9	18	0.75	10.7	13.1	0.6	0.05	40
Accessorie	111	6.55	21	2	2	4	15	2	1	0.5	NA	1	7.3	7	1.6	0.05	84
Accessorie	112	15	75	25	6	3	3	9	2	0.2	7	2.2	21.3	1.8	7.8	0.05	300
Accessorie	113	52.5	10	8	5	0	1	2	0	0.8	NA	1.1	15.6	3	15	0.05	40
Accessorie	114	21.08	313	62	13	8	16	44	3	0.9	6	0.35	5.7	3.5	8.3	0.05	1252
Accessorie	115	18.98	349	118	27	7	5	57	3	0.9	NA	0.6	1.7	13.5	10.2	0.05	1396
Accessorie	116	3.6	8	6	3	2	1	0	0	0.8	927	0.01	11.5	8.5	0.4	0.05	32
Accessorie	117	3.6	11	5	2	2	1	0	0	0.8	661	0.01	11.5	8.5	0.4	0.05	44
Accessorie	118	174.99	170	100	23	20	20	310	6	0.8	1	1.4	13.8	8.2	0.4	0.05	680
Accessorie	119	9.99	15	12	4	0	4	3	1	0.8	NA	0.4	11.1	7.6	0.5	0.05	60
Software	120	670	20	2	4	2	4	4	3	0.7	110	0.25	5.8	1.4	7.8	0.15	80
Software	121	133.08	34	15	2	2	10	5	4	0.7	NA	3.2	7.4	5.5	1.4	0.15	136
Software	122	124.99	394	187	63	42	86	55	38	0.8	1	0.15	7.6	5.5	1.2	0.2	1576

Preparing the Data

There was not that much pre-processing that needed to be done.

One thing we did was converting all the data into numbers as otherwise we won't be able to run them into the ML algorithms properly.

The only other modification we made initially was to exclude the "BestSellersRank" column as it has a missing entries.

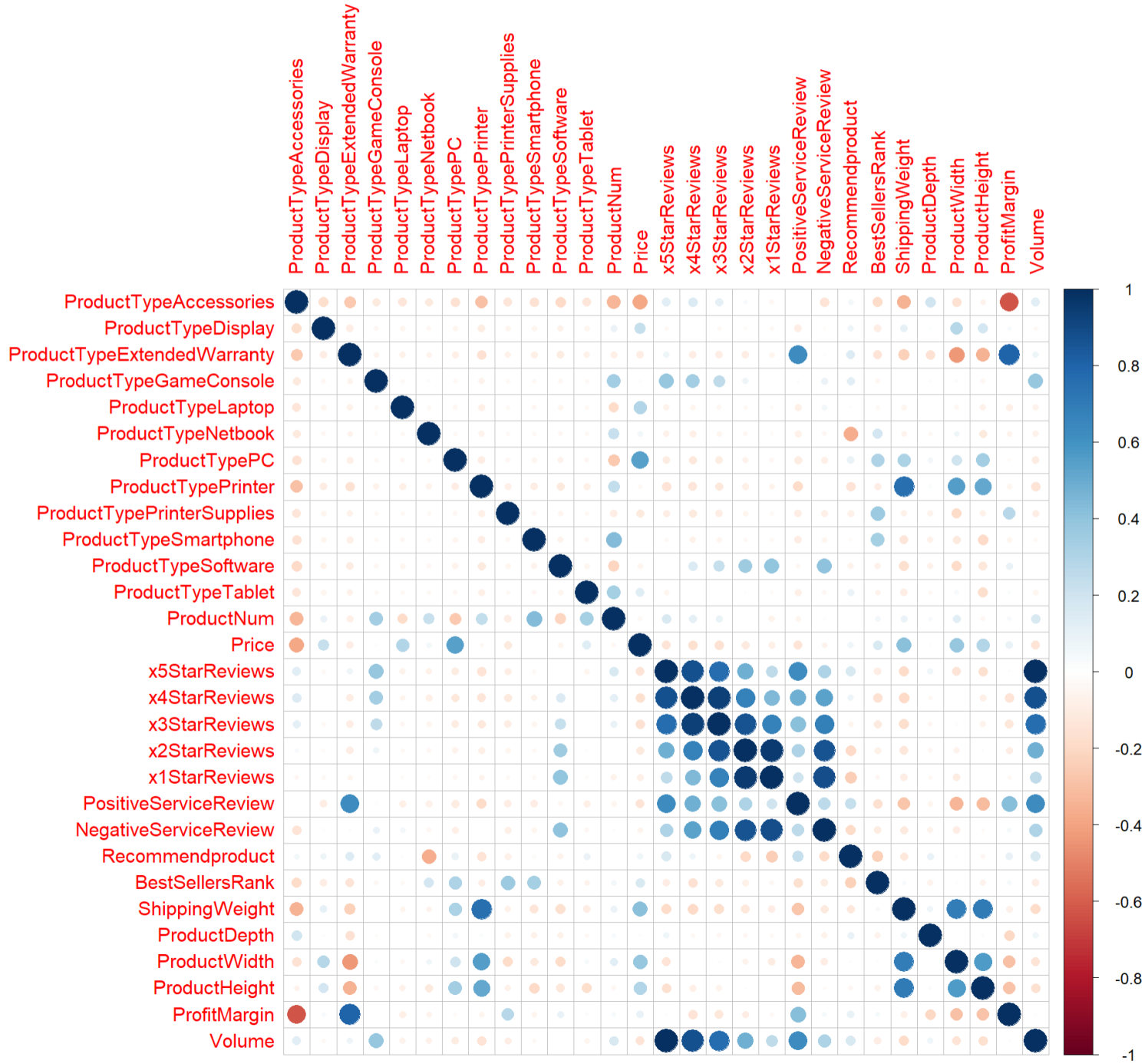
Theoretically, we could have kept this column and recode the missing entries with the average of the column, but when we did a correlation test (which we'll talk more about later) it showed that this column has little to no effect on volume. Therefore, we did the right thing by removing it.

Correlation Test

For this test, we tested to see the correlation coefficient between each attribute. (This is measured from -1 to 1 where if its closer to the end points then that means there is a strong relationship and near 0 means there is no relationship)

What we want to focus on is each attribute compared to volume to see if it is either dark red or dark blue.

We can see that the reviews seem to be the one that mostly affect sales (volume).



How do we predict the Sales?

In order to do this, the steps are:

1. Take the existing products data and split them into two sets: one for training and one for testing.
2. Pick a few machine learning models and train them with the training dataset.
3. Test each model with the other dataset (testing) and see how fairs against one another
4. Pick the model that does the best and then use that to predict the sales on the new product data.

Steps that we took with each Model

1. We took the training part of the existing data and feed them to a model.
2. We tested how well the model fairs by seeing the initial r^2 values. These can range between -1 to 1 and the closer they is to either -1 or 1, the more accurate that model is.
3. If we feel the model's r^2 values are normal, then we can do a post resample test where we see it summarizes all of them into one r^2 result. This determines the overall accuracy of the model.

Model 1: Linear Regression

The first model we tried was a simple linear regression model and it didn't perform well as the initial r^2 value ended up being an exact 1.

Hypothetically, this means that the result is perfect, however, having a result that is exactly 1 is never realistically possible. This usually means there is a larger problem at hand.

Therefore, we didn't move ahead with the post resample test with this model. (Technically speaking we didn't move ahead as, unlike other models, there was only one result for the r^2 value)

Model 2: Random Forest

- This model's initial results seemed more realistic as they ranged from 0.59 - 0.66. At this point, these aren't any great results either as r^2 near 0.60 means its mediocre accuracy.
- However, when we move on ahead to resample test, that is when we see the realistic accuracy being oof about 0.86. Which is, relatively speaking, very good.
- The one thing that surprised us is how did the accuracy increase dramatically in the resample test?

Model 3: Gradient Boosting Machine

This model's results were not that great as the initial r^2values were ranging from 0.41 – 0.60. Which means overall it fairs worse than the previous one (Random Forest).

Nevertheless, we still went ahead and did the resample test. The result ended up being about 0.44; which is more inline with the initial results. However, this is on the lower end of the spectrum and weak score overall.

Therefore, we can cross out GBM as a potential model to use for this data.

Model 4: Support Vector Machine

- The steps for this model were slightly different in that we didn't get any initial r^2 values, but instead had one final value.
- Nevertheless, the value ended up being about 0.87; which is the highest one we've seen yet.
- Therefore, we might pick this model to use to build our final predictions.

Which Model do we use?

In the end, we can pick either Random Forest (RF) model or Support Vector Machine (SVM) as both ended up having the highest accuracy from the models we tested.

For now, we decided to use SVM to predict the sales as it did have the slight edge over RF.

The predicted sales were recording in an output file called "predictions.csv"

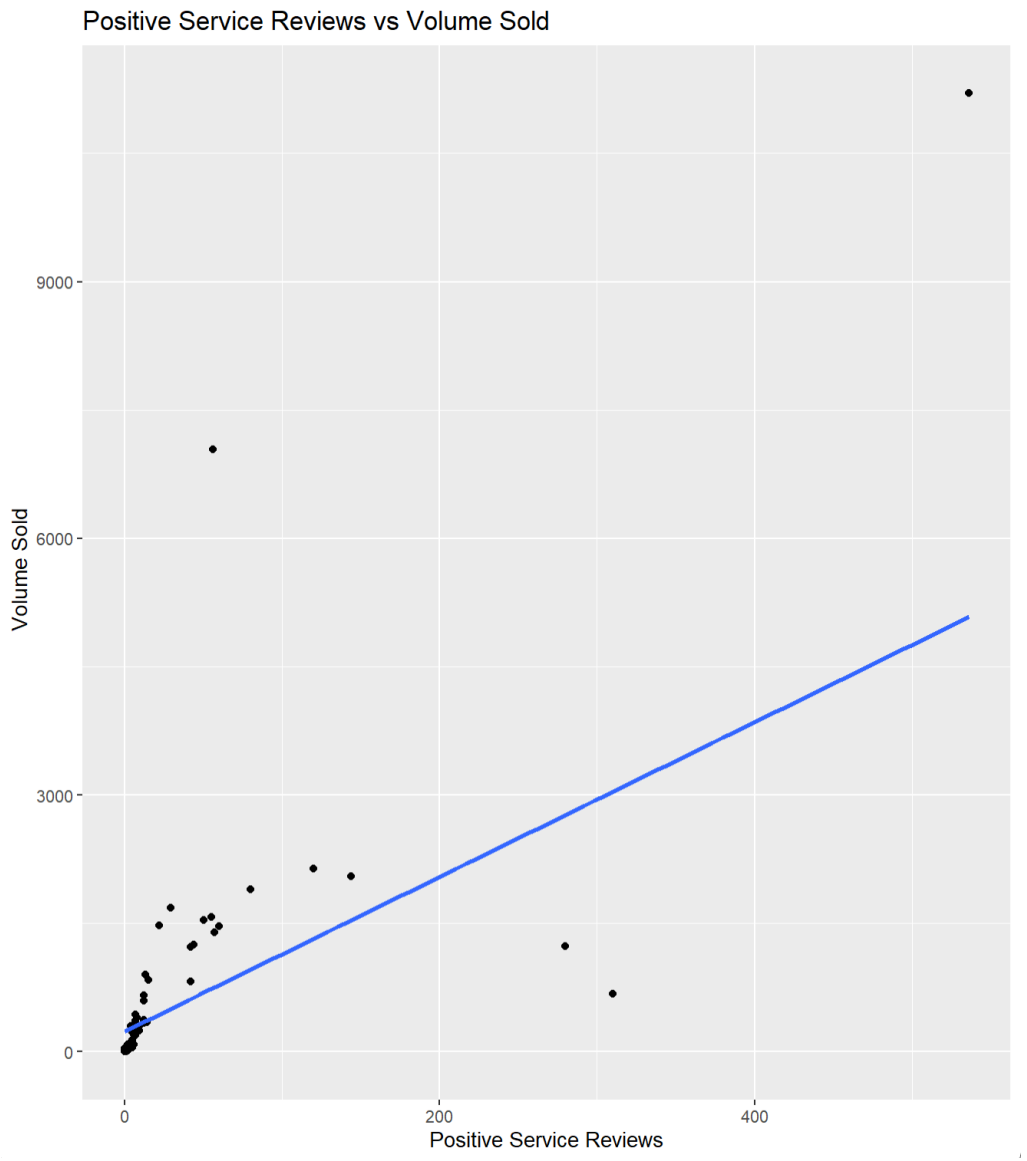
How do Positive Service Reviews affect Sales?

Since we had two kinds of service reviews, positive and negative, we looked at both separately.

Here shows a scatterplot of the positive reviews and how they affect the volume sales. We can see a relation as when the number of positive reviews increase, so does the volume sales.

We even plotted a best fit line and found the correlation coefficient; which ended up being around 0.62.

While this might be mediocre amount for a ML model accuracy score, but for a simple correlation, it is enough for us to conclude that, for now, the more positive reviews so have a positive effect on sales.

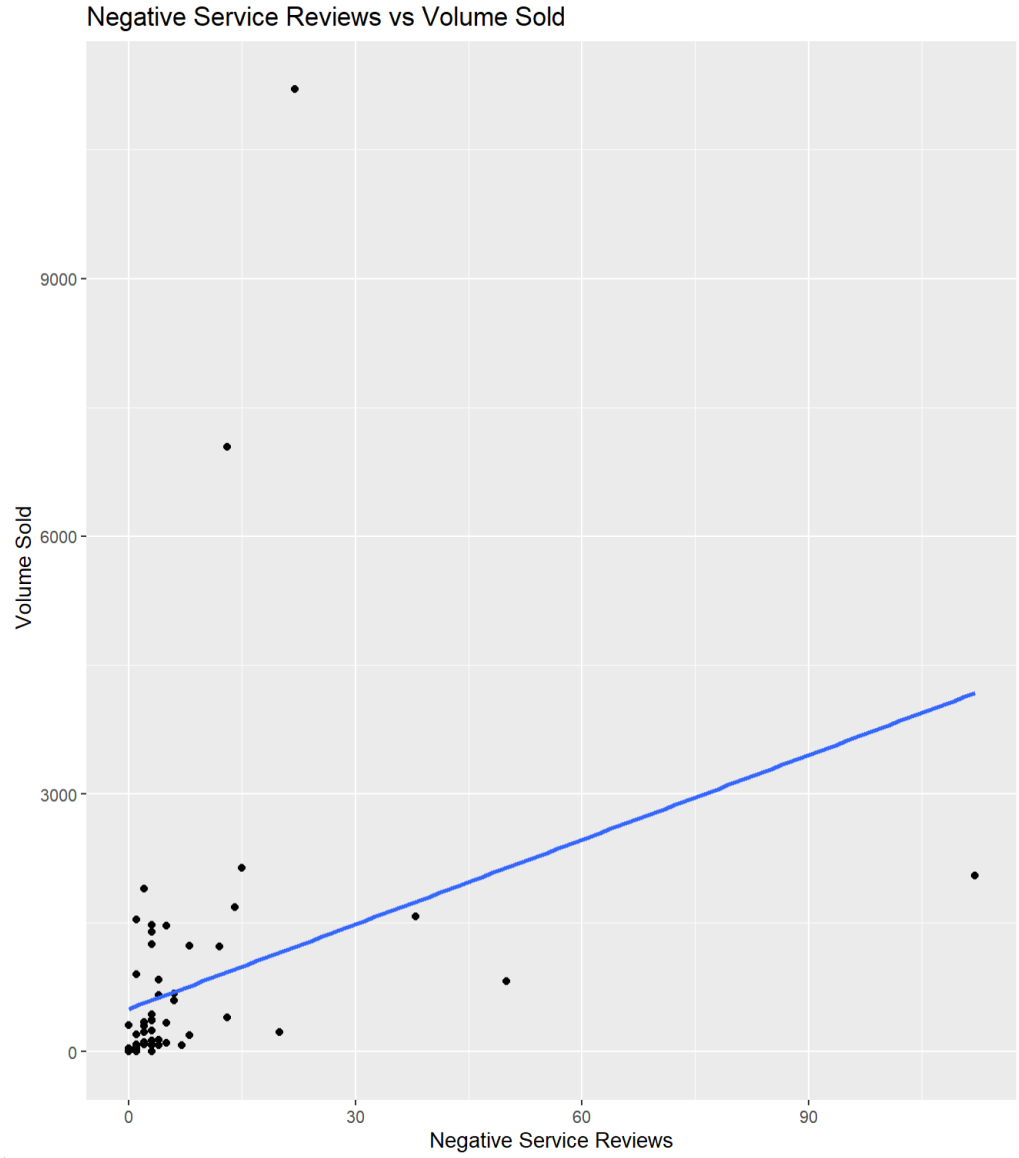


How do Negative Service Reviews affect Sales?

When it comes to the negative reviews, we expected there to be a negative relationship as logically if one item is badly reviewed, the sales should be negative.

However, looking at the scatterplot, we find that the opposite is true. There seems to be a very weak positive relationship.

The correlation of this is about 0.31 which proves that there is indeed a very weak relationship between negative reviews and sales.



Service Reviews Analysis

Why was there a positive relationship between negative reviews and volume sold?

Our best guess is that the relation might be the other way around where it's the volume sold that is affecting the negative reviews. For example, if a product is sold well, the chance of it being reviewed more is likely because more people have the product in hand. Therefore, it will logically have more negative reviews.

For positive reviews, the same rule applies, but we still feel that this might a two-way relationship where both affect one another. We say this as the correlation was much stronger here.

Conclusion

Predicting sales of four different product types: PC, Laptops, Netbooks and Smartphones:

- We were able to predict the sale by using the SVM model.

Assessing the impact services reviews and customer reviews have on sales of different product types.

- Negative reviews seem to be affected by the volume sales, meanwhile positive reviews seem to be both affect and be affected by the volume sales.