1. Task Description

Hello,

The sales team has again consulted with me with some concerns about ongoing product sales in one of our stores. Specifically, they have been tracking the sales performance of specific product types and would like us to redo our previous sales prediction analysis, but this time they'd like us to include the ‘product type’ attribute in our predictions to better understand how specific product types perform against each other. They have asked our team to analyze historical sales data and then make sales volume predictions for a list of new product types, some of which are also from a previous task. This will help the sales team better understand how types of products might impact sales across the enterprise.

I have attached historical sales data and new product data sets to this email. I would like for you to do the analysis with the goals of:

- 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

When you have completed your analysis, please submit a brief report that includes the methods you employed and your results. I would also like to see the results exported from R for each of the methods.

Tasks Deliverables - Sales Prediction Report: A report in a Zip file that includes:

- A brief summary in Word or PowerPoint of your methods and results that include:
  - The algorithms you tried.
  - The algorithm you selected to make the predictions, including a rationale for selecting the method you did and the level of confidence in the predictions.
  - Your sales predictions for four target product types found in the new product attributes data set
  - A chart that displays the impact of customer and service reviews have on sales volume.
- The results of each model you constructed, exported from R

Thanks,
Danielle

Danielle Sherman
Chief Technology Officer
Blackwell Electronics
www.blackwellelectronics.com

2. Task Solution

Two data sets (existingproductattributes2017.csv and newproductattributes2017.csv) were provided by the CTO, to conduct an analysis to predict which types of products might impact sales across the enterprise. Three models were developed to complete this task, each model contains one algorithm:

Model 1 - Support Vector Machine (SVM)
Model 2 – Random Forest
Model 3 – Gradient Boosting

2.1 Models Configuration

Model 1 – Support Vector Machine with Linear Kernel (svmLinear)

This model contains the following processes and parameters:

- Pre-processing (normalization, dummy variables)
- Data Partition (80/20)
- Train Control (method = "cv", number = 10)
- Grid = 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2,5
- Method = svmLinear
- Predictions
- Post Resample

Model 2 – Random Forrest

This model contains the following processes and parameters:

- Pre-processing (normalization, dummy variables)
- Data Partition (80/20)
- Train Control (method = "cv", number = 10, search = 'random')
- Method = rf
- Predictions
- Post Resample

Model 3 – Gradient boosting with algorithm eXtreme Gradient Bossting (xgbTree)

This model contains the following processes and parameters:

- Pre-processing (normalization, dummy variables)
- Data Partition (80/20)
- Train Control (method = "cv", number = 10)
- Method = xgbTree
- Predictions
- Post Resample

Algorithm Selection:

SVMLinear: This algorithm can be used for regression or classification problems, C is the tuned parameter

RF: This algorithm can be used for regression or classification problems, several parameters can be tuned

xgbTree: This algorithm can be used for regression or classification, several parameters can be tuned

3. Results

All models were run under the specified conditions. The model with the best performance is the Random Forest, a heat map of sales is plotted in the image below (Figure 1), the Top 5 sales product types are Game Console, Tablet, PC, Laptop and Netbook.

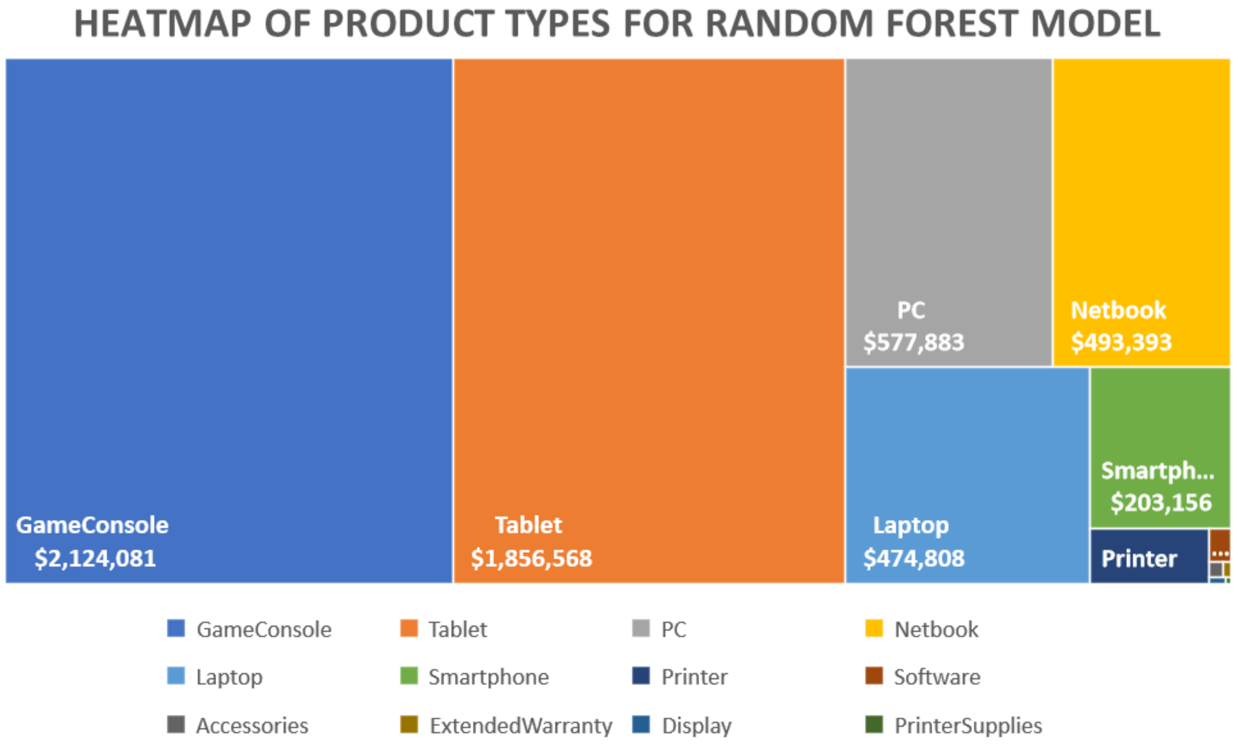


Figure 1 – Product Type Sales Forecast

The quality metrics selected for this process was the post resample (Figure 2), in this process the random forest algorithm pose the highest accuracy (R=0.99) with the lowest RMSE (49.10), between the 3 models. Therefore, predictions from random forest will be used to determine the sales impact of the different product types.

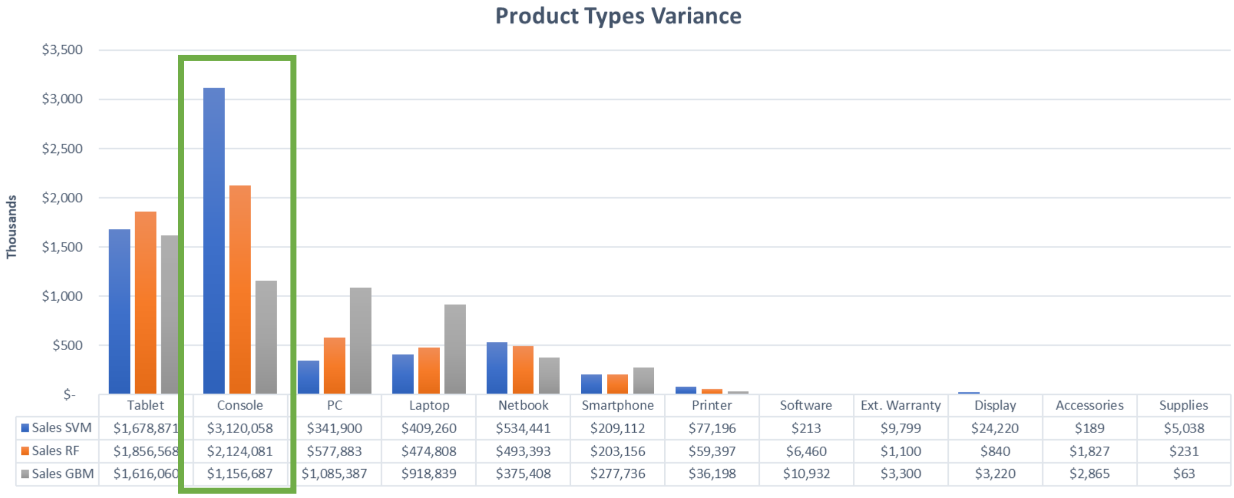
	SVMLinear	Random Forest	Gradient Boosting
RMSE	128.2034	49.1054	138.9466
R-squared	0.9599	0.9930	0.9431
MAE	111.6307	26.7462	70.4094

Figure 2 – Post Resample

In the context of this exercise a lower RMSE means a higher concentration of the data around the line of best fit. (Lower spread of the residuals); the RMSE gives a relatively high weight to large errors. As a result, lower values of RMSE indicates better fit.

In the other hand R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. In our case Random Forest showed a higher square correlation, closer to 1, which is good, cause shows a better correlation between the variables.

Another factor to be considered in this analysis is the variance in each product type. Figure 3 shows the variance on each product type for the 3 models. The highest variance is in the “Game Console” product type, the variance between Sales SVM and Sales GMB is ($1,93MM ~63%) and the variance between Sales SVM and Sales RF is ($0.995MM ~32%).



	Tablet	Console	PC	Laptop	Netbook	Smartphone	Printer	Software	Ext. Warranty	Display	Accessories	Supplies
■ Sales SVM	$1,678,871	$3,120,058	$341,900	$409,260	$534,441	$209,112	$77,196	$213	$9,799	$24,220	$189	$5,038
■ Sales RF	$1,856,568	$2,124,081	$577,883	$474,808	$493,393	$203,156	$59,397	$6,460	$1,100	$840	$1,827	$231
■ Sales GBM	$1,616,060	$1,156,687	$1,085,387	$918,839	$375,408	$277,736	$36,198	$10,932	$3,300	$3,220	$2,865	$63

Figure 3 – Variance between models and product types

The variance of sales across the models (Figure 4) shows the following results:

A - 10,53% off in sales predictions for Sales with SVM model vs Sales with RF model

B – 5,40% short in sales predictions for Sales with GBM model vs Sales with RF model

ProductType	Sales SVM		Sales RF		Sales GBM	
Tablet	$	1,678,871	$	1,856,568	$	1,616,060
Console	$	3,120,058	$	2,124,081	$	1,156,687
PC	$	341,900	$	577,883	$	1,085,387
Laptop	$	409,260	$	474,808	$	918,839
Netbook	$	534,441	$	493,393	$	375,408
Smartphone	$	209,112	$	203,156	$	277,736
Printer	$	77,196	$	59,397	$	36,198
Software	$	213	$	6,460	$	10,932
Ext. Warranty	$	9,799	$	1,100	$	3,300
Display	$	24,220	$	840	$	3,220
Accessories	$	189	$	1,827	$	2,865
Supplies	$	5,038	$	231	$	63
Grand Total	**$**	**6,410,297**	**$**	**5,799,744**	**$**	**5,486,695**
Variance	$	610,554	**Empty**		$	(313,049)
Percentage		10.53%	**Empty**			-5.40%

Figure 4 – Sales variances for Models SVM, RF and GBM

The impact of service review in the Sales RF (Figure 5) shows that product types with high positive service reviews (blue bars) has the highest sales values. While, product types with lower positive reviews (orange bars) showed lower sales values.

Additionally, the impact of customer reviews in the Sales RF (Figure 6) shows that product types with the highest reviews has the highest sales values. While, product types with lower customers reviews showed lower sales values.

ProductType	Positive Service Review	Negative Service Review	Sales Values
GameConsole	91	25	$ 2,124,081
Tablet	118	32	$ 1,856,568
PC	19	8	$ 577,883
Netbook	36	36	$ 493,393
Laptop	13	7	$ 474,808
Smartphone	31	20	$ 203,156
Printer	5	1	$ 59,397
Software	4	2	$ 6,460
Accessories	4	1	$ 1,827
ExtendedWarranty	0	3	$ 1,100
Display	1	1	$ 840
PrinterSupplies	1	0	$ 231

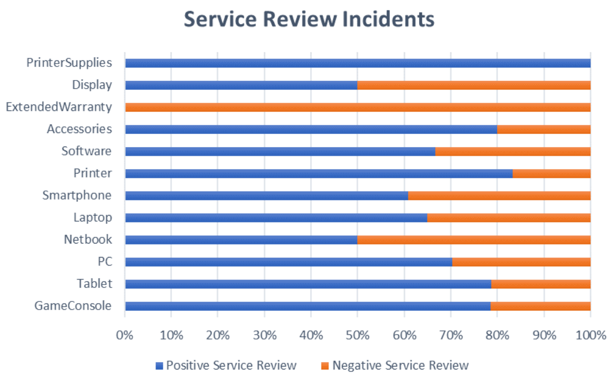


Figure 5 – Impact of service reviews in sales forecast

ProductType	x1StarReviews	x2StarReviews	x3StarReviews	x4StarReviews	x5StarReviews	Sum Reviews	Sales RF
GameConsole	103	73	124	349	1987	2636	$ 2,124,081
Tablet	283	181	254	503	1239	2460	$ 1,856,568
PC	46	24	24	37	147	278	$ 577,883
Netbook	75	55	39	142	357	668	$ 493,393
Laptop	12	7	5	13	82	119	$ 474,808
Smartphone	114	73	66	79	291	623	$ 203,156
Printer	3	1	3	8	88	103	$ 59,397
Software	8	1	3	18	29	59	$ 6,460
Accessories	15	4	7	3	55	84	$ 1,827
ExtendedWarranty	1	1	1	1	0	4	$ 1,100
Display	2	0	0	0	4	6	$ 840
PrinterSupplies	0	0	0	0	5	5	$ 231
	662	**420**	**526**	**1153**	**4284**	**7045**	**$ 5,799,744**

Figure 6 – Impact of customer reviews in sales forecast for Model RF

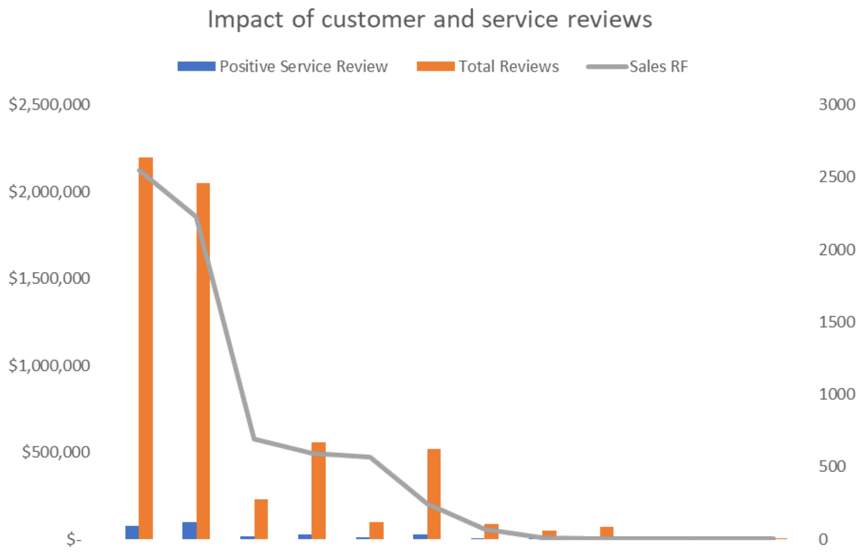


Figure 7 – Impact of customer and service reviews in sales forecast for Model RF

Figure 7 shows a positive trend in sales for product types with high customer and service reviews, while low values of customer and service review represents lower sales.

4. Results Discussion

The sales predictions with the model RF show a variance of +11% and -5% in sales with models SVM and GBM respectively, depending of the business context a range between of +/- 10% could be acceptable.

Product types with high customer and positive service reviews shows a positive trend in sales, while product with lower values in these reviews shows a negative trend in sales.

5. Recommendations

Use the model RF to conduct the predictions since the R-squared values were the highest one and RMSE was the lowest one.

6. R scripts

Provided in a zip file