Report

Raul Vazquez, Gerardo Parrello

December 14, 2018

Executive Summary

What is this report about?

Our purpose is to predict profitability for a list of new products based on the performance and characteristics of current products, including in our prediction the product category for each one to asses the impact of specific attributes, customer and service reviews, for certain product types.

Preamble

From previous analysis, we know that 4 Star Reviews, 2 Star Reviews and Positive Service Reviews are the best predictors of Sales Volume. In this review, we have added Product Type as a possible predictor of Sales Volume.

Conclusions

After our analysis we conclude that Product Type is not a good predictor of Sales Volume, although that a product pertains to the type 'Game Console' is significant.

We can also say that there is not enough data to asses the impact of customer and service reviews for each product category independently, but we can assert their importance in overall.

Finally, we have included a file named *new.csv* with the predicted Sales Volume and Total Profit using our best model.

Recommendations

Since we do not posses enough data to asses significance of variables for each product category, we recommend regrouping the data by a different categorization which creates bigger but still realistic groups.

Technical Analysis

Preprocessing

In order to perform modeling over our dataset, we have applied the following transformations:

- Outlier elimination: we have excluded two observations that present Volume over 7000
- Normalization of numerical variables, with exception of the target attribute, using z-transformation
- Feauture selection

Feature Selection

In order to assert the importance of variables in respect to Sales Volume and select the most significant, we have run a Linear Regression model:

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -980.76 -108.94   -8.55   38.58 1245.38
##
## Coefficients: (1 not defined because of singularities)
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    386.697    239.842   1.612  0.11389
## ProductType.Accessories        137.224    251.417   0.546  0.58790
## ProductType.Display             67.129    291.156   0.231  0.81870
## ProductType.ExtendedWarranty   225.726    318.078   0.710  0.48158
## ProductType.GameConsole        938.180    419.250   2.238  0.03023 *
## ProductType.Laptop             -42.510    306.990  -0.138  0.89048
## ProductType.Netbook            -39.100    413.901  -0.094  0.92516
## ProductType.PC                 -45.104    309.672  -0.146  0.88485
## ProductType.Printer            -19.768    263.954  -0.075  0.94063
## ProductType.PrinterSupplies    -35.542    309.858  -0.115  0.90919
## ProductType.Smartphone         354.828    307.064   1.156  0.25396
## ProductType.Software            -9.929    290.960  -0.034  0.97293
## ProductType.Tablet                  NA         NA      NA       NA
## x4StarReviews                  341.808     81.176   4.211  0.00012 ***
## x2StarReviews                    9.323     64.927   0.144  0.88647
## PositiveServiceReview          218.718     73.440   2.978  0.00466 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 335.8 on 45 degrees of freedom
## Multiple R-squared:  0.7658, Adjusted R-squared:  0.6929
## F-statistic: 10.51 on 14 and 45 DF,  p-value: 6.051e-10
```

From this we reassert our previous conclusion about the significance of 4 Star Reviews and Positive Service Reviews. Nonetheless we find that 2 Star Reviews becomes not significant when introducing Product Type (as dummies), whereas pertaining to category 'Game Console' is now significant.

Modeling

To predict Sales Volume, we have trained the following models:

- Random Forest
- Support Vector Machines with Linear Kernel
- Support Vector Machines with Radial Basis Function Kernel
- Support Vector Machines with Polynomial Kernel
- Linear Regression
- k-Nearest Neighbors
- Stochastic Gradient Boosting

For performance metrics we have used repeated cross validation of 5 repeats over 10 folds. For optimization we have used a tune lenght of 5.

Performance Metrics

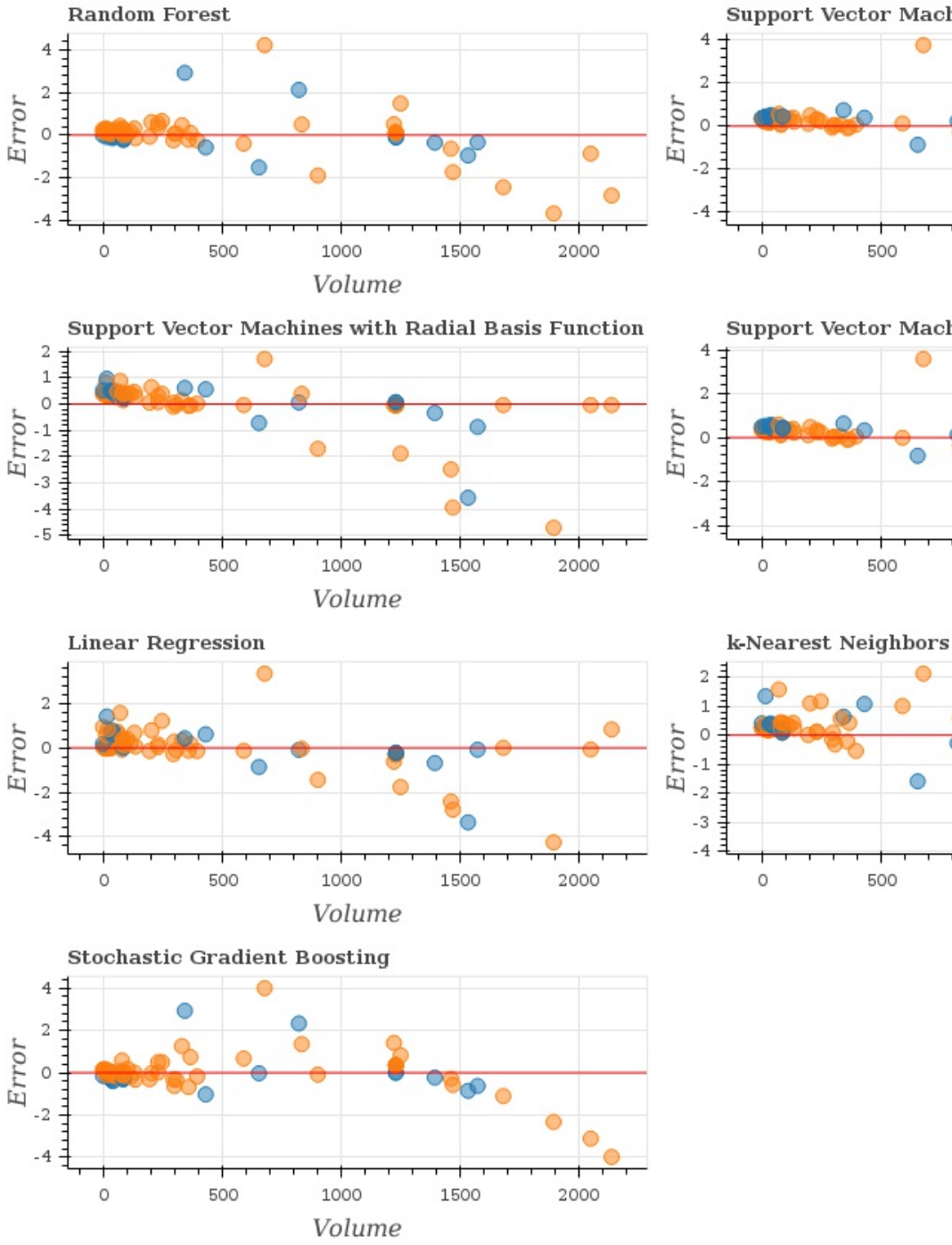
Using our performance metrics, we have determined our best model to be Random Forest, as we can see below. Columns marked with * refer to the performance of our models over our testing set; without, over the training set.

	Rsquared	Rsquared*	MAE	MAE*	RMSE	RMSE*
Random Forest	0.93	0.88	103.96	106.55	177.68	206.98

	Rsquared	Rsquared*	MAE	MAE*	RMSE	RMSE*
Stochastic Gradient Boosting	0.91	0.84	140.82	125.62	212.52	254.76
Support Vector Machines with Linear Kernel	0.74	0.79	261.70	134.72	439.45	280.59
Support Vector Machines with Radial Basis Function Kernel	0.68	0.78	283.82	156.20	456.09	295.89
Linear Regression	0.67	0.77	278.78	178.16	460.49	286.96
k-Nearest Neighbors	0.67	0.86	251.32	135.61	391.03	236.52
Support Vector Machines with Polynomial Kernel	0.51	0.78	274868.33	149.10	687374.52	288.67

Normalized Error Plots

Below we can see normalized (z-transformed) prediction error over volume for all our models. In orange, observations pertaining to



our training set; in blue, to our testing set.