Blackwell Electronics
Predicted Profit for Potential New Products

By: Rhys Hewer, Jessica Krueger, and Luca Vehbiu

Executive Summary

Goal: This report contains an analysis of previous Blackwell sales data to predict the profitability of potential new products within specific product categories. Those categories include Laptop, Netbook, PC, and Smartphone.

Process: A Random Forest regression model (RMSE = 158.635, R-squared = .720) based on 183 examples and 17 attributes was used to predict the sales volume of each new potential product. We removed 62 examples (25% of the total dataset) from the analysis because those examples were either outliers or they included information that we believed to be false/inconsistent. Due to the presence of NA values, we imputed 493 values in the data set using a kNN method.

Conclusion: Our profitability predictions are included below for each of the four product types.

Product_type	Product_ID	Prices	Volume_Pred	Profit_margin	Profit_Pred
Laptop	176	$1,999.00	287.5923667	0.23	$132,226.34
Laptop	175	$1,199.00	302.5902	0.15	$ 54,420.85
Laptop	173	$1,199.00	370.2165667	0.1	$ 44,388.97
Netbook	180	$ 329.00	1236.566133	0.09	$ 36,614.72
Netbook	181	$ 439.00	141.9598333	0.11	$ 6,855.24
Netbook	178	$ 399.99	103.8376	0.08	$ 3,322.72
Netbook	183	$ 330.00	45.9703	0.09	$ 1,365.32
PC	171	$ 699.00	408.8553333	0.25	$ 71,447.47
PC	172	$ 860.00	226.5147	0.2	$ 38,960.53
Smartphone	193	$ 199.00	365.0122333	0.11	$ 7,990.12
Smartphone	196	$ 300.00	129.6927667	0.11	$ 4,279.86
Smartphone	194	$ 49.00	391.2352	0.12	$ 2,300.46
Smartphone	195	$ 149.00	89.9517	0.15	$ 2,010.42

Cautions: The data provided for this analysis contained several inconsistencies and a large amount of NA values. Data of this nature is generally considered unreliable. It is difficult to produce a reliable predictive model based on unreliable data, and therefore we cannot be as confident in these predictions as we have been in previous predictions. The predictions that we produced are still far better than attempting to guess volume values for potential new products, however, there is a higher likelihood for error in these predictions than there has been in our previous predictions.

Blackwell Electronics
Predicted Profit for Potential New Products
By: Rhys Hewer, Jessica Krueger, and Luca Vehbiu

Model Selection

Model Selected: Random Forest

Random Forest Testing Metrics:

```
       RMSE    Rsquared         MAE
158.6354882   0.7199089  96.9975076
```

Random Forest Model Details:

```
Random Forest

139 samples
 17 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 124, 127, 126, 124, 125, 125, ...
Resampling results across tuning parameters:

  mtry  RMSE      Rsquared   MAE
  17    215.3605  0.6326557  131.4088
  23    215.4991  0.6331136  131.5321
  27    215.2996  0.6320929  131.3742
  33    215.3133  0.6320343  131.2579

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 27.
```

Models Tested but Not Selected: GBM & SVM Linear

GBM Testing Metrics:

```
       RMSE    Rsquared         MAE
266.2508587   0.3150793 186.2382445
```

GBM Model Details:

```
Stochastic Gradient Boosting

139 samples
 17 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 125, 125, 125, 125, 125, 127, ...
Resampling results across tuning parameters:

  interaction.depth  n.trees  RMSE      Rsquared   MAE
  1                   50      292.6389  0.5258805  195.8995
```

```
  1                  100      285.5820  0.5434452  195.1091
  1                  150      282.8464  0.5517778  194.5217
  2                   50      285.1712  0.5423041  197.0385
  2                  100      274.0722  0.5497542  187.5091
  2                  150      272.3969  0.5486831  189.8861
  3                   50      291.1214  0.5192336  199.6744
  3                  100      285.2798  0.5173164  198.1335
  3                  150      279.3610  0.5198382  195.4526

Tuning parameter 'shrinkage' was held constant at a value of 0.1

Tuning parameter 'n.minobsinnode' was held constant at a value of 10
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were n.trees = 150, interaction.depth
 = 2, shrinkage = 0.1 and n.minobsinnode = 10.
```

GBM Predictions included negative values. Negative values are not possible for sales volume:

```
summary(testPredGBM)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-222.80   44.03  147.83  201.10  258.03 1262.01
> summary(testing$Volume)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   19.75   63.00  181.86  205.00 1472.00
```

SVM Linear Testing Metrics:

```
       RMSE    Rsquared         MAE
189.0792954   0.6025204  95.4475480
```

SVM Linear Model Details:

```
Support Vector Machines with Linear Kernel

139 samples
 17 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 125, 125, 125, 125, 125, 127, ...
Resampling results:

  RMSE      Rsquared   MAE
  220.1338  0.5873777  118.7563

Tuning parameter 'C' was held constant at a value of 1
```

SVM Linear predictions included negative values. Negative values are not possible for sales volume:

Blackwell Electronics
Predicted Profit for Potential New Products

By: Rhys Hewer, Jessica Krueger, and Luca Vehbiu

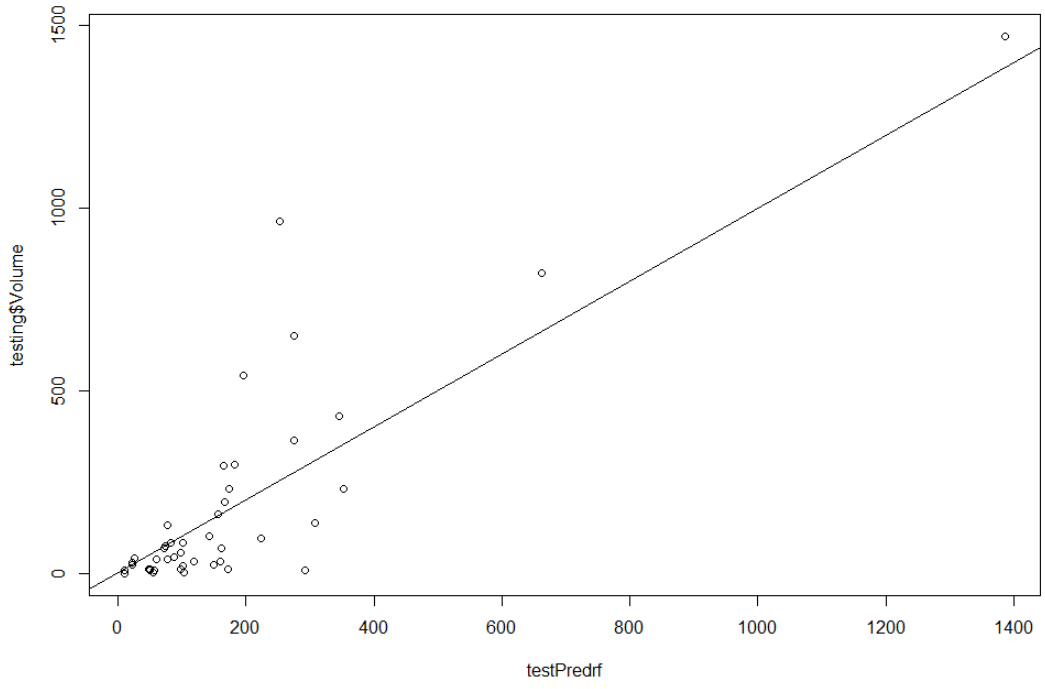
```
summary(testPredSVM)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-118.67   40.28   81.98  154.16  146.57 1424.74
> summary(testing$Volume)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   19.75   63.00  181.86  205.00 1472.00
```

Model Selection Rationale:

We focused our efforts on the Random Forest model because it was the only model to produce non-negative predictions for sales volume. Sales volume cannot be a negative number, so any model that produces a negative number is inherently inaccurate.

The Random Forest model also produced the best testing metrics (RMSE = 158.635, R-squared = .720), but we are concerned that the training metrics (RMSE = 215.300, R-squared = .632) are slightly different than the testing metrics. We believe that the poor quality of the data might be the reason for this.

The model overall performs well, but as you can see from the below plot of the predicted volume vs. the actual volume in the testing set, there are some significant errors. This model is certainly better than attempting to guess sales volume for new products, however, due to the poor quality of the data, we cannot be as confident in these predictions as in previous predictions.



Blackwell Electronics
Predicted Profit for Potential New Products
By: Rhys Hewer, Jessica Krueger, and Luca Vehbiu

Data Cleaning Performed

Data cleaning steps were performed on the data resulting in the removal of 62 examples from the analysis. A total of 183 out of 245 original examples were used in the final model. The removed examples account for 25% of the total examples.

Imputation: The data set included 493 NA values in the star review and service review columns. We chose not to remove the NA's because doing so would dramatically reduce the size of our data set and make analysis more difficult. Instead, we chose to impute the NA values using the kNN method (k = 3).

The data set also included 30 examples with zero values for product weight and dimensions. This seemed illogical since a physical product must have dimensions. Of the 30 products with zero values, 24 of them were printer supply products. However, there were only 26 total printer supply products in the data set. There was not enough available data to impute the zero values with a reasonable value, and we could not leave the values as zero knowing that they were false. Therefore, we chose to remove these examples from the data set.

For the other 6 products with zero values for dimensions and weight, we imputed the zero values using the median value of each attribute by product type.

Outliers: This data set included a significant amount of outliers. However, five outliers were deemed to be extreme outliers and were removed from the analysis. The extreme outliers were:

Product #	Outlier Type	Outlier Value
150	Volume	11,204
198	Volume	7,036
118	Positive service review	310
102	Price	$2,249.99
123	Negative service review	112

Inconsistent Examples: We removed 30 examples from the Extended Warranty product category because this product type represents a service and not a product. Since our goal is to predict the sales volume of products, we should only include products in the analysis.

We also removed 3 examples (product ID's 166, 129, and 363) because each example included values for star and service reviews but each had a zero value for sales volume. It seems illogical that an item can have reviews but no sales, therefore we determined that this data is unreliable and we removed it.

Feature Selection

In order to select features, we used several methods including producing a correlation matrix (see page 7), a decision tree, and several scatter plots to understand which features have the most effect on the sales volume. Based on those methods, we selected the below attributes to create each model:

- Product type

- 5-star reviews
- 4-star reviews
- 2-star reviews
- Positive service reviews
- Negative service reviews
- Price

Attributes Not Used: In order to reduce noise and create a model that performs well, we excluded irrelevant and redundant attributes.

Irrelevant attributes are those that have low or no correlation with sales volume and/or are not likely to impact sales volume. The following attributes were deemed irrelevant:

- Product depth
- Product width
- Product height
- Product weight
- Profit margin
- Gender
- Age
- In-store
- Relative price
- Environmental impact
- Durability standard

The following attributes were deemed redundant because they were colinear with other attributes or were providing similar information:

- 3-star reviews
- 1-star reviews
- Would consumer recommend product

Blackwell Electronics
Predicted Profit for Potential New Products

By: Rhys Hewer, Jessica Krueger, and Luca Vehbiu

Correlation Matrix