**Understanding Customers and Predicting Profitability**

**Introduction**

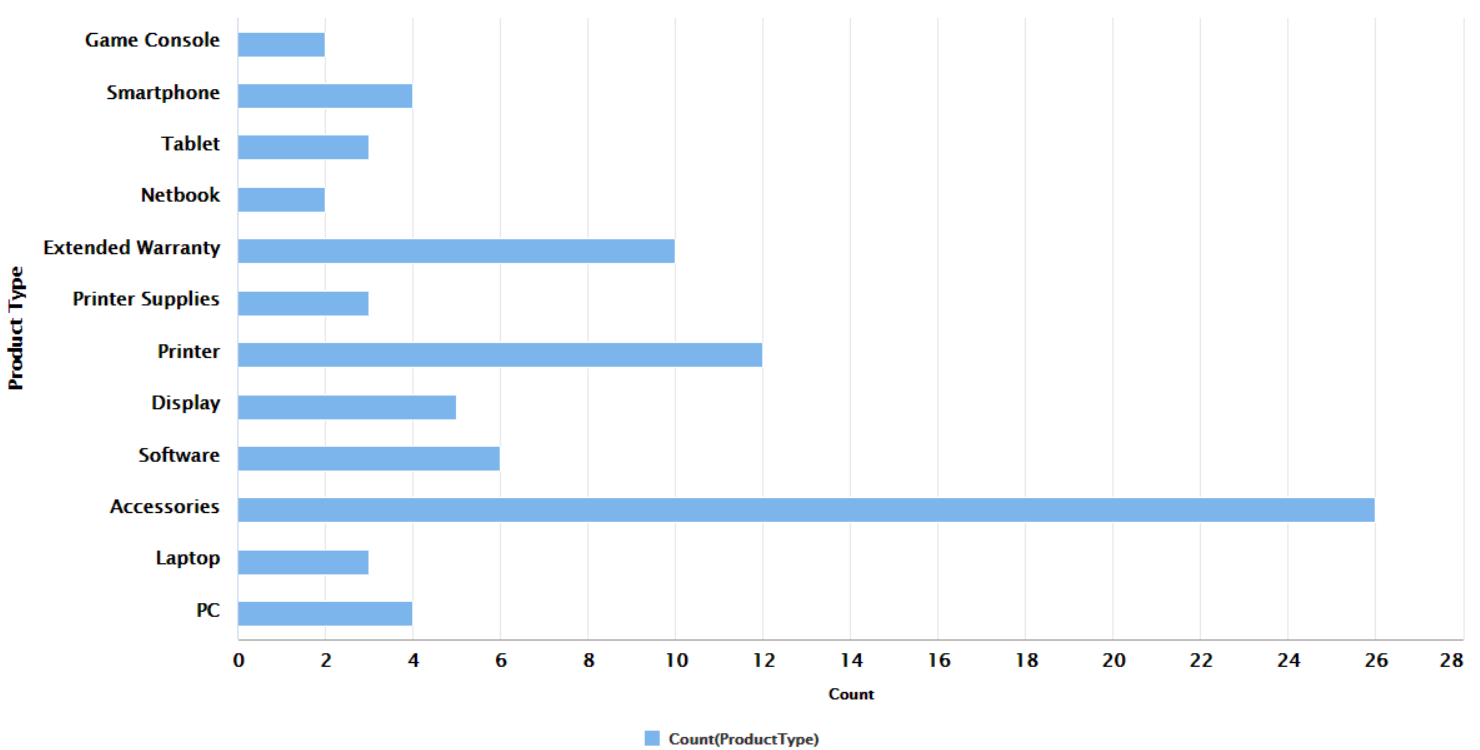
Blackwell – electronics retails is considering to add new products to their inventory. 17 new products have been shortlisted taking into consideration Blackwell’s business strategy. Further, the CTO of the company wants to help sales item in predicting profitability of the potential new products and short listing top 5 products.

The main purpose of this task is to investigate sales volume from which profitability of all new products can be estimated by performing detailed regression analysis using RapidMiner. Two regression analysis methods k-NN (k Nearest Neighbor) and SVM (Support Vector Machine) have been used to perform this investigation. GBT (Gradient Boosting Trees) is also used in this analysis to improve decision trees. For this study these models have been tested and trained on 80 existing products data.

After several iterations of adjusting parameters for both k-NN (parameter k), SVM (parameter C and kernel type), and GBT (no. of trees and learning rate); SVM resulted to be best fit model with least Root Mean Square Error (RMSE) = 330.102 +/- 202.226 (micro average: 381.802 +/- 0.000) and R2 (Squared Coefficient) = 0.872 +/- 0.152 (micro average: 0.951). This model was then applied to all the potential 17 new products to predict their sales volumes and estimate profitability.

**Method: Cleaning, Preprocessing and Analyzing Data**

Figure 1 is a bar graph showing number of products in each product type category. A total of 80 existing products have been analyzed in this study.



**Figure 1: No. (count) of existing products in each product type**

***Data Cleaning***

Data provided for analysis includes Product type, product number, price, profit margin, sales volume, shipping weight, product height/depth/width, best sellers rank, # of customer reviews (stars 5 to 1)/recommendations, and positive/negative service reviews. Initially the data was cleaned by the removing any spaces or special characters in attribute labels. Also missing values in Best sellers rank was replaced by “NA”. Also, “x” was placed in any attribute label that had numerals as initial characters.

***Pre-Processing***

Product Type (polynomial attribute) and Best Sellers Rank (contains missing data) attributes were removed from existing data since they cannot be measured by regression analysis. Product numbers is an integer attribute and is a type of unique data, therefore has been set as special attribute “ID” in RapidMiner. Since sales volume is the dependent variable (because that is what we need to investigate for new products) it was set as “label” in Rapidminer.

* **Existing Data Statistics:**

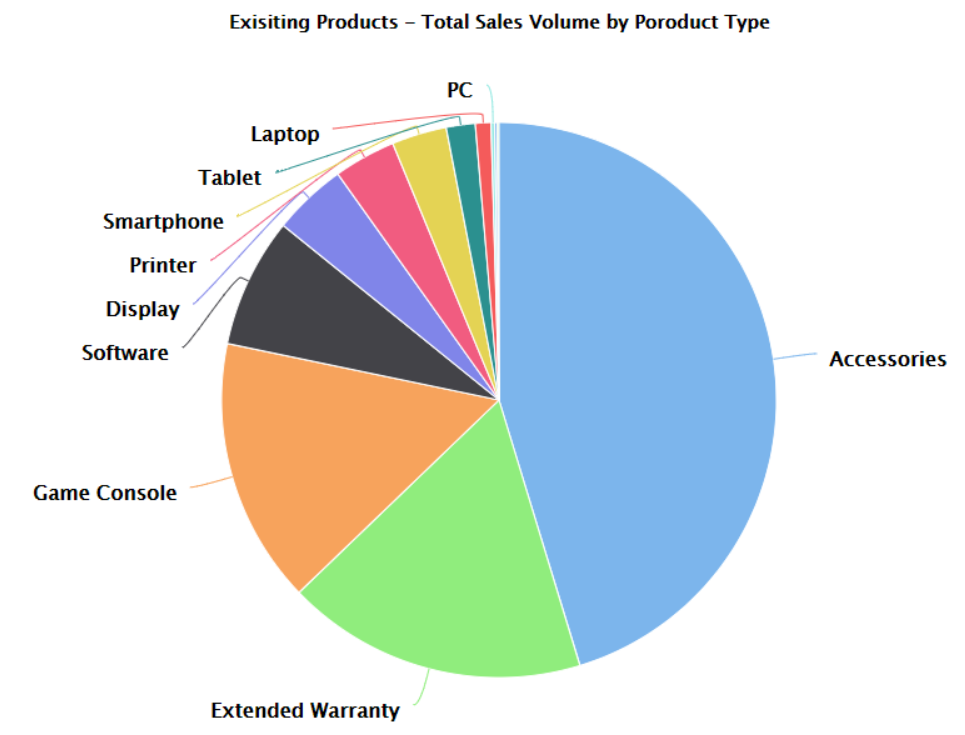
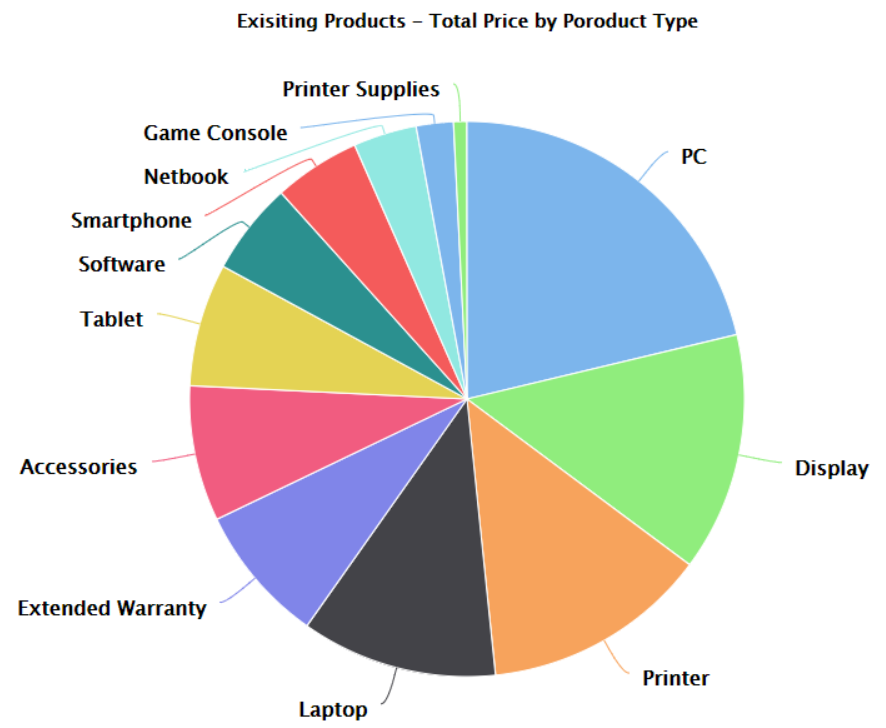
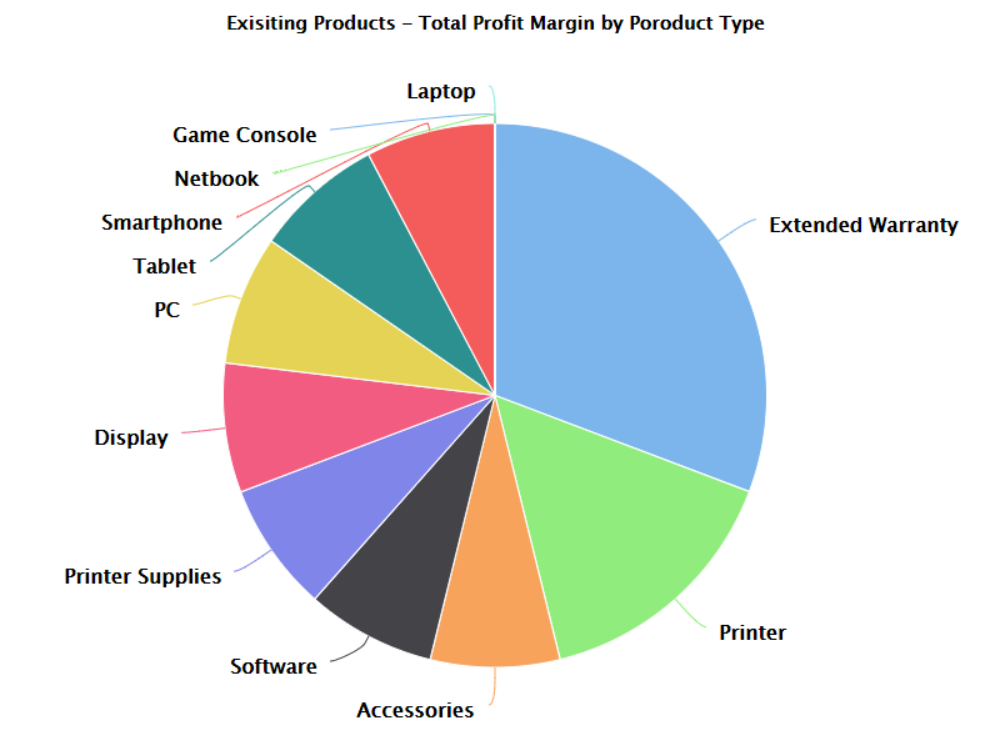
Figure 2, 3 and 4 show total sales volume, total profit margin and total prices, respectively for each of the product type in existing products.

Figure 2: As per the pie chart showing total sales volume for each product type accessories, extended warranty, game console, software and display are top 5 ranking products. All other products have fairly low sales volume as compared to top 5 products.

Figure 3: As per the pie chart showing total profit margin for each product type extended warranty, printer, accessories, software and printer supplies are top 5 ranking products. It can also be interpreted from pie chart that accessories, software, printer supplies, display, PC, tablet, smartphone, netbook, and game console have fairly similar profit margin.

**Figure 2: Existing products Total Sales Volume by Product type**

Figure 4: As per the pie chart showing total profit margin for each product type PC, display, printers, laptop and extended warranty are top 5 ranking products. Extended warranty and accessories, and tablet have similar prices.

**Figure 4: Existing products Total Price by Product type**

**Figure 3: Existing products Total Profit Margin by Product type**

Table 1: Show summary of top 5 products based on total sales volume, total profit margin and total price. Extended warranty is found to be in all three attributes. Accessories and Software were in top 5 products in total sales volume and total profit margin. Printer was in top 5 products in total profit margin and total price, while display was in top 5 products in total sales volume and total price. PC and Printer supplies were in top 5 products in total price and total profit margin, respectively



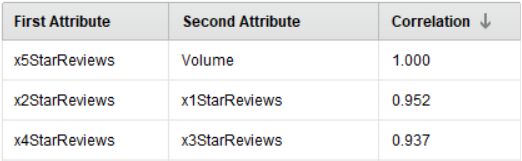
**Table 1: Top 5 existing products based on total sales volume, profit margin, and price**

* **Normalization**

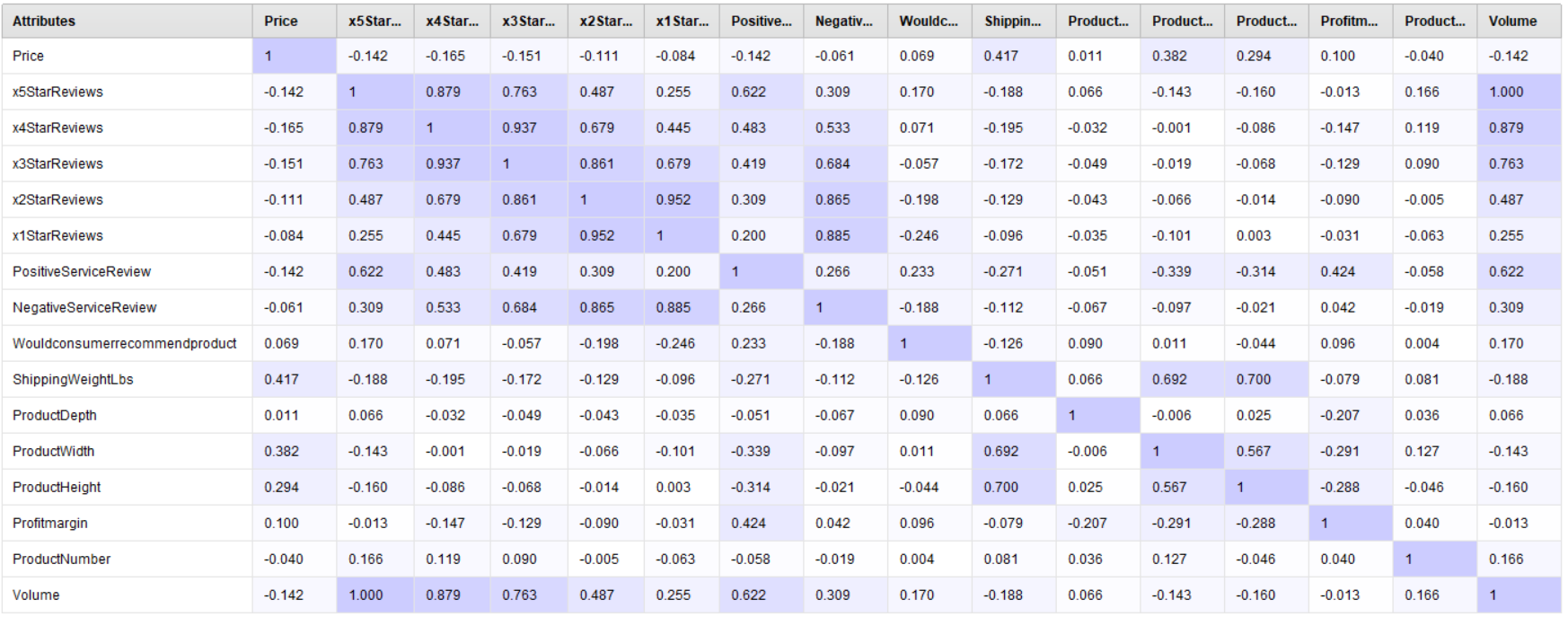
Some of the attributes in data have different units for example prices are in dollar, and weight in lbs. It is important to normalize these attributes and bring them on same scale when dealing with regression analysis in machine learning. Normalization of data can therefore help in preventing any possible bias in the model. For this particular task price, product shipping weight, 5/4/3/2/1 customer reviews, customer service feedback (positive and negative), product height, width, depth, and profit margin were normalized on scale of 0 to 1 using Range transformation. Since volume (dependent variable) and product number (ID) are special attributes they have not been normalized.

* **Feature Selection using Correlation Matrix**

Correlation matrix was used to understand relationship between data. This matrix shows strength of the relationships between every feature with values ranging somewhere between -1 to 1. Positive numbers indicate positive association and negative numbers indicate inverse relationship. In this task the correlation matrix showed that dependent variable volume has positive correlation (value = 1.0) with 5 star reviews (Table 2). From Table 2 (pairwise table) two attributes 1 star review and 3 star reviews were found with correlation matrix higher than 0.90 (0.95 and 0.937 respectively) and therefore were removed from dataset to address collinearity. Table 3 shows correlation matrix, showing how each attribute is correlating with other attributes within the data set.



**Table 2: Pairwise table showing correlating attributes with correlation greater than 0.9**



**Table 3: Correlation Matrix**

***Analyzing Data***

* **Training, Optimizing and Selecting Best Fit Model**

Three algorithms k-NN (k Nearest Neighbor), SVM (Support Vector Machine) and GBT (Gradient Boosting Trees) were trained and assessed for this machine learning regression analysis. Root Mean Square Error (RMSE) and Squared Correlation (R2) metrics were used as performance measures for all three algorithms.

*Root Mean Square Error*: Measures difference between values predicted by model and the observed values

*Squared Correlation (R2):* Measure of variance i.e., to what extent variance in one variable explains variance in second variable. It can measure anywhere from 0 to 1. Acceptable range is between 0.8 -0.9.

**K-NN (k Nearest Neighbor)** Measures the distance between new observations and the nearest *k* training instances to make predictions. It is based on feature similarity. In regression problem the output value is average or median of the values of it k- nearest neighbors. Current dataset was trained and tests for k = 1, 2, 3, 4, and 5. Results for RMSE and R2 metrics are shown in Table 4.

**Table 4: k-NN Performance Metrics**

**SVM (Support Vector Machine)** SVM is also a technique that can be used in both classification and regression problems. SVM maintains all the main features that characterize the algorithm. This algorithm tries to fit error within a certain threshold. Kernel Type (mathematical method of mapping dataset into higher dimensional space where data is easily separable to make better predictions) and C – Complexity Constant (sets the tolerance for misclassification) are the two parameters tuned for optimization of this model.

Too large C 🡪 Overfitting

Too low C values 🡪 Underfitting

Table 5 shows several iterations of SVM-regression model with different C values and Kernel Type DOT. C = 54 yields RMSE = 330.102 +/- 202.226 (micro average: 381.802 +/- 0.000) and R2 = 0.872 +/- 0.152 (micro average: 0.951).



**Table 5: SVM Performance Metrics**

**GBT (Gradient Boosted Tree Algorithm)** GBT is a machine learning technique used to predict models for regression and classification problems. It generates prediction model typically in form of decision trees (ensemble of weak predictive models). It is a forward learning ensemble method that obtains predictive results through gradually improved elements. GBT involves 3 specific elements:

1. Loss function to be optimized (Loss function is a measure indicating how good model coefficients are fitting underlying data)
2. Weak learner (decision trees) to make predictions
3. Additive model to add weak learner to minimize loss function

Two GBT parameters learning rate and no. of trees have been tuned for optimizing this model. Smaller the learning rates better the model. Lower learning rates require more iteration. For the purpose of this task GBT algorithm was trained and tested on several iterations of number of trees and Learning Rate = 0.1. Results for RMSE and R2 metrics are shown in Table 6.



**Table 6: GBT Performance Metrics**

After optimizing all three regression models and comparing performance metrics tables 4, 5, and 6 it can been interpreted that Support Vector Machine (SVM) with C (complexity constant) = 54 and Kernel Type “DOT” yields least level of comparative error ***RMSE = 330.102 +/- 202.226 and*** ***R2  = 0.872 +/- 0.152.*** Therefore, SVM has been choosen and used as best fit model to predict sales volume for 17 new products.

**Predicting Sales Volume and Estimating Profitability of New Products**

After selection this model, new product data was cleaned and preprocessed in same manner as existing products. Followed by applying the best fit model to predict sales volume and estimate profitability of the 17 new products.

***Profitability = estimated volume × profit margin × price***

Table 7 shows sales volume and estimated profitability for 17 new products ranked from highest to lowest based on profitability. From Table 7 – product numbers 171, 187, 172, 176 and 186 are the top 5 products with highest profitability. Three product types PC, Laptop, and Tablets have been identified as products with highest profitability.



**Table 7: New Products – Predicted Sales Volume and Estimated Profitability. Ranked Highest to Lowest**

**Conclusion**

The purpose of this study was to use regression analysis on existing products to predict sales volume and estimate profitability of potential new products. Furthermore, this study was conducted to help sales team with short listing top 5 new products out of 17 new potential products that can be added to Blackwell’s inventory. After detailed regression analysis of data using k-NN, SVM and GBT algorithm showed that SVM is the best model with least error and high correlation. Using this model the top 5 new potential products that can be added to Blackwell’s inventory are PC, Tablet and Laptop.