* A ranking of potential new products, ordered from highest to lowest profitability
  + Which product has the highest number of 5 Star Reviews?
    - Product# 150, 2801 5-star reviews, Accessories
  + Which product has the highest sales volume?
    - Product# 150, 11204 Sales Volume, Accessories
  + Do the data types match your expectations?
    - BestSellersRank: Integer: Yes
    - NegativeServiceReview: Integer: Yes
    - PositiveServiceReview: Integer: Yes
    - Price: Real: Yes
    - ProductDepth: Real: Yes
    - ProductHeight: Real: Yes
    - ProductNumber: Integer: No
      * This is like an id #. It doesn’t mean anything more.
      * Use Set Role operator to set as ID
    - ProductType: Polynomial: Yes
    - ProductWidth: Real: Yes
    - ProfitMargin: Real: Yes
    - SalesVolume: Integer: Yes
    - ShippingWeightsLbs: Real: Yes
    - WouldConsumerRecommend: Real: Maybe
      * What do these values mean?
      * Are they percentage of consumers that would recommend? If so, then real is appropriate.
      * Are they on a scale from 1-10 but in decimal form? If so, then ordinal
    - x1Star: Integer: Yes
    - x2Star: Integer: Yes
    - x3Star: Integer: Yes
    - x4Star: Integer: Yes
    - x5Star: Integer: Yes
  + Use Statistics and Charts to look for other relationships that may be useful to your analysis
    - Interesting relationships:
      * Idea: Certain ProudctTypes sell more than others
        + ProductType frequency distribution (histogram)

ProductType vs Sales Volume (of existing products)

* + - * + Product Type , Sales Volume vs BestSellersRank
        + Product, Sales Volume vs WouldCustomersRecommend
        + Product, Sales Volume vs x5Star
        + Product, Sales Volume vs x5Star, x4Star
        + ProductType, Sales Volume vs Height,Width
        + ProductType, Sales Volume vs Price
    - Correlation Matrix
      * Relationships where correlations are above .90 or -.90:
        + 5-star and SalesVolume: 1.00
        + 4-star and 3-star: 0.937
        + 2-star and 1-star: 0.952
        + Variables removed: 1-5 stars,
* A summary of performance metrics from each individual classifier you ran
  + **KNN**
    - KNN when K=5
      * root\_mean\_squared\_error: 1060.938 +/- 1124.059 (micro average: 1504.243 +/- 0.000)
      * squared\_correlation: 0.346 +/- 0.393 (micro average: 0.060)
    - KNN when K=2
      * root\_mean\_squared\_error: 1301.736 +/- 1251.028 (micro average: 1761.556 +/- 0.000)
      * squared\_correlation: 0.371 +/- 0.308 (micro average: 0.021)
    - KNN when K=1
      * root\_mean\_squared\_error: 1297.100 +/- 1098.671 (micro average: 1663.983 +/- 0.000)
      * squared\_correlation: 0.000 +/- 0.000 (micro average: 0.000)
  + SVM
    - SVM when C=.5
      * root\_mean\_squared\_error: 1100.136 +/- 1069.167 (micro average: 1496.364 +/- 0.000)
      * squared\_correlation: 0.387 +/- 0.289 (micro average: 0.025)
    - SVM when C=.1
      * root\_mean\_squared\_error: 1128.743 +/- 1064.838 (micro average: 1514.778 +/- 0.000)
      * squared\_correlation: 0.379 +/- 0.296 (micro average: 0.000)
    - SVM when C=.3
      * root\_mean\_squared\_error: 1154.518 +/- 1048.574 (micro average: 1523.965 +/- 0.000)
      * squared\_correlation: 0.382 +/- 0.292 (micro average: 0.001)
    - SVM when C=.9
      * root\_mean\_squared\_error: 1099.014 +/- 1094.731 (micro average: 1512.093 +/- 0.000)
      * squared\_correlation: 0.399 +/- 0.288 (micro average: 0.020)
  + Gradient Boosted Trees
    - Gradient Boosted Trees when Trees = 100
      * root\_mean\_squared\_error: 823.135 +/- 1095.513 (micro average: 1325.777 +/- 0.000)
      * squared\_correlation: 0.756 +/- 0.205 (micro average: 0.243)
    - Gradient Boosted Trees when Trees = 200
      * root\_mean\_squared\_error: 842.307 +/- 1046.051 (micro average: 1301.646 +/- 0.000)
      * squared\_correlation: 0.778 +/- 0.201 (micro average: 0.254)
    - Gradient Boosted Trees when Trees = 1000
      * root\_mean\_squared\_error: 1023.531 +/- 997.526 (micro average: 1393.975 +/- 0.000)
      * squared\_correlation: 0.723 +/- 0.182 (micro average: 0.200)
* A brief summary of the optimized model you selected and your rationale for selecting it. Include the parameter settings for this model.
  + I chose the Gradient Boosted Model with 100 trees because it has the lowest root-mean-squared error and the second highest squared\_correlation of all the different models and different parameters.
* Products Ranked by Profitability

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ProductType** | **ProductNumber** | **BrandName** | **SalesVolume** | **PredictedProfit** |
| Tablet | 187 | Amazon | 9,813.72 | $390,586 |
| PC | 171 | Dell | 1,132.23 | $197,858 |
| Laptop | 173 | Apple | 1,586.63 | $190,237 |
| Netbook | 180 | Acer | 5,356.13 | $158,595 |
| PC | 172 | Dell | 830.27 | $142,806 |
| Netbook | 181 | Asus | 2,733.44 | $131,998 |
| Tablet | 186 | Apple | 1,862.11 | $117,127 |
| Game Console | 199 | Sony | 4,149.86 | $93,368 |
| Smartphone | 193 | Motorola | 973.13 | $21,302 |
| Smartphone | 194 | Samsung | 970.56 | $5,707 |
| Laptop | 175 | Toshiba | 5.32 | $957 |
| Netbook | 178 | HP | 24.61 | $788 |
| Netbook | 183 | Samsung | 24.52 | $728 |
| Monitor | 201 | Asus | 52.88 | $370 |
| Smartphone | 196 | Motorola | -12.92 | -$426 |
| Smartphone | 195 | HTC | -19.55 | -$437 |
| Laptop | 176 | Razer | -4.04 | -$1,858 |