Assignment 3: Model Development Using Watson Analytics

Daanish Ahmed
DATA 610 9080
Fall 2016
dsahmed2334@yahoo.com
Professor Vahe Heboyan
November 13, 2016

Introduction

An increasing number of organizations today are using predictive analytics to aid their decision-making, and one of the most widely-used models for this process is the decision tree. This algorithm is used to partition a dataset into segments based on certain classifications or questions, resulting in a tree-like structure that is easy to understand (Berson, Smith, & Thearling, n.d.). In this assignment, my goal is to implement decision trees in order to better understand their impact towards decision making. Using IBM's Watson Analytics, I will analyze two datasets—one with a categorical target variable and one with a continuous one. The results of my analysis will provide insight on the factors that affect the target variables in each case. In the end, these findings should also provide implications for solving real-life problems in my organization.

The two datasets in my analysis were obtained from IBM's list of sample datasets, with one set focusing on telecommunications customer churn and the other based on sales profits and costs (Stacker, 2015). Both of these datasets are very robust—the customer churn dataset features 7043 cases and 21 possible input values (see Table 1 in Appendix A), while the sales and marketing dataset contains 24,743 cases and 14 possible inputs (see Table 2 in Appendix A). The dataset on customer churn has several notable inputs such as total and monthly charges, online security, internet service, tenure, tech support, and online backup. Some of the variables that can be used to develop predictive models include churn, payment method, and total charges. For the dataset on sales, the possible inputs include revenue, planned revenue, product type, quantity, product cost, and retailer country. Here, predictive models may be developed from gross profit, quantity, and revenue. In this assignment, I will analyze decision trees generated from one target in each dataset. For the customer churn dataset, I will focus on churn—a categorical variable. For the sales dataset, the output in question is gross profit, which is a continuous variable. But before I begin my analysis of the decision tree models, I will briefly address the steps that I took to prepare the datasets for analysis in Watson Analytics.

Data Cleansing and Preparation

I analyzed both of the datasets in Microsoft Excel to ensure that they were cleansed of outliers, missing values, and errors. For the customer churn dataset, my calculations yielded that no outliers existed in any of the inputs. However, 11 missing values were found for the total charges variable. Although decision trees are not significantly impacted by missing data (Kane, 2015), I chose to remove these cases so that the dataset would be better suited for any additional analysis. The loss of information in this case is minimal since there are over 7000 cases. For the sales dataset, I found that the data needed significant cleansing. There were initially over 84,000 cases, but 70% of those cases were missing all numeric information such as product cost, revenue, and gross profit. As a result, these entries contributed little at all to my analysis, and removing them would greatly improve the quality of the results. Even after doing this, the dataset was still extremely robust with 24,743 cases. One additional problem was that hundreds of upper outliers existed in columns like revenue, product cost, quantity, and gross profit. I truncated all of these outliers to a value three standard deviations from the mean. The loss of data here would also be minimal; though some columns had over 700 outliers, this number is relatively negligible for a dataset of over 24,000 entries.

Customer Churn Model

Now that the data cleansing is complete, I will begin to discuss my findings from the customer churn dataset. For this categorical model, a distribution of the two possible outcomes reveals that the customer churn rate is around 27% (see Figure 2 in Appendix A). Looking at the decision tree shows a breakdown of which categories of customers are most likely to experience churn (see Figure 1 in Appendix A). Some of the strongest predictors in this model include total charges, monthly charges, internet service, online security, contract, tenure, and online backup. One observation from this decision tree is that having a shorter contract period leads to a higher customer churn. The churn rates for a month-to-month contract are higher than those for a one or two year contract, indicating that there is less commitment among customers with shorter contracts. Regarding internet service, those without internet have the lowest churn rates, while those with fiber optic service have higher churn than those with DSL. This suggests that customers without service tend to have far fewer technical problems compared to those with service. And although fiber optics is faster and more reliable than DSL, the fact that it lacks availability in many areas ("ADSL vs Cable," 2013) might be a reason why churn is higher for fiber optics. Additionally, customers without tech support are more likely to experience churn than those without it. By using a spiral visualization, I found that the single best predictor is tenure; having a higher tenure makes churn far less likely. Also, the best combination of variables is total charges and internet service. The churn rate goes down as total charges increase, which suggests that those who spend more money will be less likely to churn as long as they are satisfied with their services. It may suggest that some customers do not care as much about internet speed or price, but rather availability and ease of use.

After looking at the decision tree itself, some additional insights can be gained from examining the top five decision rules for this model (see Figure 3 in Appendix A). One thing to note is that all five of these rules have month-to-month contracts and fiber optic internet services. These results are therefore consistent with my findings on contract period and internet service. Likewise, all but one of these rules has the lowest tenure, reflecting that those with lower tenure are more likely to withdraw from the company's services. Many of these cases also use multiple lines, suggesting that having multiple lines can lead to higher churn. These decision rules are very helpful and can offer great benefits to my organization. According to Bahnsen, Aouada, and Ottersten (2015), if companies can identify which customers are most likely to churn, then they will know the segment of their market that they need to appeal to in order to prevent loss. Like this dataset, my organization also offers internet and telecommunications services. Though not all of our products or services are identical, there are enough similarities for my organization to consider learning from these rules in order to prevent churn. Later into this paper, I will go into details regarding specific applications within my organization.

Gross Profit Model

After looking at the customer churn model, I will now analyze the decision tree for gross profit to see how certain inputs affect a continuous variable. This tree puts great emphasis on the strength of the factors influencing gross profit (see Figure 5 in Appendix A). My observations suggest that some of the strongest predictors are revenue, planned revenue, product type, unit sale price, and quantity. The single best predictor is revenue—having a higher revenue will of

course lead to stronger profits. But in addition to this, certain product types generate a higher profit as well. In this case, putters, sleeping bags, packs, and watches have the greatest demand and therefore generate the highest gross earnings. Furthermore, having a larger quantity of products will generally lead to a higher gross profit—especially if those particular products sell well. In several cases, a product's retailer country can also directly affect the gross earnings of that product. The expenses from transferring merchandise made in different countries may lower the total profit earned from those sales (Cheptea, Emlinger, & Latouche, 2013). Additionally, the gross profit across certain years can indicate whether or not a product was in high demand during those years. Overall, the best combination of variables is revenue (or planned revenue) and product type. Using a heatmap (see Figure 6 in Appendix A), I noticed that when a product has a high expected revenue and is of a highly-demanded product type, it results in a much greater gross profit than if the product only met one of these conditions.

When looking at the top five decision rules (see Figure 7 in Appendix A), there are several insights that can be gained. Firstly, all products listed under these rules—including sleeping bags, tents, and watches—are products with the highest demand. Likewise, all of these rules apply to products with revenues exceeding $250,000. These rules are therefore consistent with earlier findings that revenue and product type are important towards increasing gross profit. Here, we also see a clear example of how the year of sale impacts the profit. The decision rules with the first and third highest gross profits are both of the same product type, quantity, revenue, and retailer country. However, the products sold in 2006 and 2007 made $63,904 more than those sold in 2004 and 2005. These results clearly provide useful insights, and my organization can learn a lot from these decision rules. Although my organization does not offer the same products as this dataset, many of the concepts are still relevant. For instance, my company can benefit from using a model like this to study a product's performance across the years in order to determine if that product is becoming more or less popular. Also, one noteworthy aspect of these findings is that a product's cost does not significantly impact the gross profit as long as that product is in demand. This suggests that my organization should focus on maintaining a large quantity of the most highly-demanded products, including the more expensive ones.

Development of Additional Models

The two models that I analyzed so far are very detailed, but I chose to refine them further to produce more meaningful results. For the customer churn model, I tried to make it more accurate by removing inputs that had a lesser impact on the results. Some of these variables include dependents, device protection, paperless billing, partners, and payment method. Afterwards, I simplified the model further by clipping branches which yielded lower churn, so that the model would focus mainly on customers who were most likely to leave the company. I clipped branches for those with one and two year contracts, those with no internet service, and those with DSL service (see Figure 4 in Appendix A). Doing this resulted in some rules being slightly different while still being located in the same general area of the decision tree. The tree became much simpler and more useful for my organization due to its focus on high customer churn areas which need most of the company's attention. The consequence is that the predictive strength is slightly weaker (by 0.4%), and the results are likely to be less accurate due to the omission of some columns. Likewise, for the gross profit model I removed weaker inputs such as product cost, unit cost, product line, and order method. I also clipped all branches where the

revenue was less than $250,000 in order to focus on the nodes with high gross profit (see Figure 8 in Appendix A). But unlike for the customer churn model, the top 5 decision rules for this model were not affected by omitting some of the variables. In the end, this tree became much more focused because it emphasized the products which would most likely result in higher profit. The change in predictive strength is even lower for this model (it decreased by 0.1%), and any loss in data due to the filtering of columns is also minimal.

Applications

Using the insights from my analysis, I will now take a deeper look at some specific applications within my organization. Regarding customer churn, one major problem is to identify the customers that are most likely to churn in order to appeal to them and ensure repeated business (Bahnsen et al., 2015). To solve this problem, a rules-based approach can be implemented in which the organization will focus its attention on the customer churn decision rules (see Figure 3 in Appendix A) to identify the customer groups that need the most attention. Yet while this approach is supported by data, the consequence is that the predictions may be imperfect. According to Bahnsen et al. (2015), there is a possibility that a customer identified as a "churner" might not churn from the company, while someone identified as a "non-churner" may end up churning. The authors claim that misclassification of these customers can result in wasted resources and a loss of some customers. However, no predictive model is 100% perfect and I would facilitate acceptance of this approach by arguing that the benefits of this approach will far surpass its shortcomings. A predictive model cannot replace human judgment, but it can still reveal prominent trends that would have been difficult to identify otherwise. Similarly for the gross profit model, I would recommend using the decision rules (see Figure 7 in Appendix A) to identify the products that would contribute the most towards boosting my organization's profit. Likewise, this strategy may also be imperfect because it may cause some potentially useful products to be neglected. But overall, the data is certain to reflect some of the biggest trends for my organization to focus on, and any neglected product categories should be countered by the increase in gross profit from using this rules-based approach.

Conclusion

In the aftermath of my analysis, I found that decision trees are indeed extremely useful for my organization when performing predictive analytics. The visualizations offered a clear connection between the input variables and the output, regardless of whether the target was a categorical variable or a continuous one. Afterwards, the creation of additional models resulted in decision trees which were more refined and specific. Though these new models were simplified from their original forms, they still retained the most important pieces of information and were easier to understand. And finally, the decision rules generated by Watson Analytics proved to be of incredible value towards solving real-life problems in my organization. While the models may not be perfect, they nevertheless provided realistic and attainable solutions that may have been difficult to develop without the use of analytics. Altogether, this data-driven approach can be invaluable for any organization's decision-making process. And as this technology grows, more businesses will find that this approach can bring them closer to success.

References

ADSL vs Cable vs Fiber Optic: What's the Difference? (2013, May 17). Retrieved November 13, 2016, from http://diallog.com/adsl-vs-cable-vs-fiber-optic-whats-the-difference/

Bahnsen, A. C., Aouada, D., & Ottersten, B. (2015, June 12). A novel cost-sensitive framework for customer churn predictive modeling. *Decision Analytics, 2*(1), 1-15. Retrieved November 10, 2016, from UMUC Library.

Berson, A., Smith, S., & Thearling, K. (n.d.). *An overview of data mining techniques* [section 2.2]. Retrieved from http://www.thearling.com/text/dmtechniques/dmtechniques.htm

Cheptea, A., Emlinger, C., & Latouche, K. (2013, April). Multinational Retailers and Home Country Food Exports. *American Journal of Agricultural Economics, 97*(1), 159-179. Retrieved from BASE.

Kane, D. (Performer) (Jan 23rd, 2015). *Data Science Part IV: Decision Trees and Random Forests* [Web]. Retrieved from https://www.youtube.com/watch?v=OByOgGXq76A

Stacker, M., IV. (2015, April 2). Guide to Sample Data Sets. Retrieved October 25, 2016, from https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets//

Appendix A

PhoneService	Churn	Contract	Dependents	DeviceProtect…	InternetService	MultipleLines	OnlineBackup	OnlineSecurity	PaperlessBilling	Partner	PaymentMethod	Se
Yes	Yes	Month-to-mo…	No	Yes	Fiber optic	Yes	Yes	No	Yes	Yes	Bank transfe…	0
Yes	Yes	Month-to-mo…	No	No	DSL	Yes	No	No	Yes	No	Electronic ch…	0
Yes	No	Month-to-mo…	No	Yes	Fiber optic	Yes	No	Yes	Yes	No	Credit card (…	0
Yes	No	Month-to-mo…	No	No	Fiber optic	Yes	Yes	No	No	No	Bank transfe…	0
Yes	Yes	Month-to-mo…	No	No	DSL	Yes	No	Yes	Yes	Yes	Electronic ch…	0
Yes	No	Two year	No	Yes	DSL	Yes	Yes	Yes	No	No	Bank transfe…	0
Yes	No	Month-to-mo…	Yes	Yes	Fiber optic	Yes	Yes	No	Yes	Yes	Electronic ch…	0
Yes	No	Two year	Yes	Yes	DSL	Yes	Yes	Yes	No	Yes	Electronic ch…	0
Yes	Yes	Month-to-mo…	Yes	Yes	DSL	No	No	No	No	Yes	Electronic ch…	0
Yes	No	Two year	No	No internet s…	No	Yes	No internet s…	No internet s…	Yes	No	Credit card (…	0
Yes	No	One year	Yes	No	DSL	No	No	Yes	Yes	Yes	Bank transfe…	0
Yes	No	One year	Yes	No internet s…	No	No	No internet s…	No internet s…	No	Yes	Bank transfe…	0
Yes	No	Month-to-mo…	No	No	Fiber optic	No	No	No	No	No	Electronic ch…	0
Yes	No	Two year	Yes	No internet s…	No	Yes	No internet s…	No internet s…	Yes	Yes	Mailed check	0
Yes	No	One year	No	Yes	Fiber optic	Yes	Yes	No	Yes	Yes	Bank transfe…	1
Yes	No	Month-to-mo…	No	No	Fiber optic	Yes	Yes	Yes	No	No	Mailed check	0
Yes	Yes	Month-to-mo…	No	No	Fiber optic	No	No	Yes	Yes	No	Electronic ch…	0
Yes	Yes	Month-to-mo…	No	No	DSL	Yes	Yes	Yes	Yes	No	Credit card (…	0
Yes	No	Two year	Yes	No	Fiber optic	Yes	Yes	No	No	Yes	Credit card (…	0
Yes	No	One year	Yes	No internet s…	No	No	No internet s…	No internet s…	No	No	Bank transfe…	0
Yes	No	Month-to-mo…	No	No internet s…	No	No	No internet s…	No internet s…	No	No	Mailed check	0
Yes	Yes	Month-to-mo…	No	No	Fiber optic	No	Yes	Yes	Yes	No	Credit card (…	1
Yes	No	Month-to-mo…	No	Yes	Fiber optic	Yes	No	No	Yes	No	Electronic ch…	1
Yes	No	One year	No	No	DSL	No	No	No	No	No	Credit card (…	0

Table 1. Telecommunications Customer Churn Dataset.

Product line	Product type	Product	Order method …	Retailer country	Year	Quantity	Gross profit	Planned revenue	Revenue	Unit price	Unit sale price	Pr
Camping Eq…	Lanterns	Firefly 2	Mail	Canada	2005	1171.0	$11,779.27	$32,050.27	$31,299.84	$27.37	$26.73	
Camping Eq…	Lanterns	Firefly 2	Mail	Japan	2005	22.0	$276.98	$602.14	$602.14	$27.37	$27.37	
Camping Eq…	Lanterns	Firefly 2	Mail	Netherlands	2005	902.0	$9,252.90	$24,687.74	$24,213.64	$27.37	$27.00	
Camping Eq…	Lanterns	Firefly 2	Mail	Germany	2005	368.0	$4,098.76	$10,072.16	$9,951.71	$27.37	$27.19	
Camping Eq…	Lanterns	Firefly 2	E-mail	Canada	2005	3343.0	$33,088.46	$91,497.91	$88,774.69	$27.37	$26.62	
Camping Eq…	Lanterns	Firefly 2	E-mail	Japan	2005	3477.0	$35,494.07	$95,165.49	$93,298.79	$27.37	$26.92	
Camping Eq…	Lanterns	Firefly 2	E-mail	Sweden	2005	23.0	$289.57	$629.51	$629.51	$27.37	$27.37	
Camping Eq…	Lanterns	Firefly 2	E-mail	Germany	2005	3703.0	$37,357.39	$101,351.11	$98,916.30	$27.37	$26.82	
Camping Eq…	Lanterns	Firefly 2	E-mail	Italy	2005	773.0	$7,845.95	$21,157.01	$20,731.86	$27.37	$26.82	
Camping Eq…	Lanterns	Firefly 2	Fax	France	2005	590.0	$5,988.50	$16,148.30	$15,823.80	$27.37	$26.82	
Camping Eq…	Lanterns	Firefly 4	Telephone	United States	2005	2781.0	$28,713.14	$81,872.64	$76,002.74	$29.44	$27.73	
Camping Eq…	Lanterns	Firefly 4	Telephone	Japan	2005	213.0	$2,311.05	$6,270.72	$6,145.05	$29.44	$28.85	
Camping Eq…	Lanterns	Firefly 4	Telephone	Korea	2005	603.0	$5,894.85	$17,752.32	$15,377.25	$29.44	$26.20	
Camping Eq…	Lanterns	Firefly 4	Telephone	China	2005	641.0	$6,954.85	$18,871.04	$18,492.85	$29.44	$28.85	
Camping Eq…	Lanterns	Firefly 4	Telephone	Singapore	2005	482.0	$5,229.70	$14,190.08	$13,905.70	$29.44	$28.85	
Camping Eq…	Lanterns	Firefly 4	Telephone	Australia	2005	805.0	$8,435.05	$23,699.20	$22,291.45	$29.44	$27.97	
Camping Eq…	Lanterns	Firefly 4	Telephone	Netherlands	2005	831.0	$8,581.15	$24,464.64	$22,617.55	$29.44	$27.53	
Camping Eq…	Lanterns	Firefly 4	Telephone	France	2005	1138.0	$11,700.21	$33,502.72	$30,690.21	$29.44	$27.42	
Camping Eq…	Lanterns	Firefly 4	Telephone	United Kingdom	2005	106.0	$1,150.10	$3,120.64	$3,058.10	$29.44	$28.85	
Camping Eq…	Lanterns	Firefly 4	Telephone	Austria	2005	631.0	$6,846.35	$18,576.64	$18,204.35	$29.44	$28.85	
Camping Eq…	Lanterns	Firefly 4	Sales visit	United States	2005	513.0	$5,566.05	$15,102.72	$14,800.05	$29.44	$28.85	
Camping Eq…	Lanterns	Firefly 4	Sales visit	Mexico	2005	2117.0	$22,323.45	$62,324.48	$59,061.45	$29.44	$28.19	
Camping Eq…	Lanterns	Firefly 4	Sales visit	Korea	2005	330.0	$3,580.50	$9,715.20	$9,520.50	$29.44	$28.85	
Camping Eq…	Lanterns	Firefly 4	Sales visit	Australia	2005	563.0	$6,108.55	$16,574.72	$16,242.55	$29.44	$28.85	

Table 2. Retail, Sales, Marketing Profit-Cost Dataset.

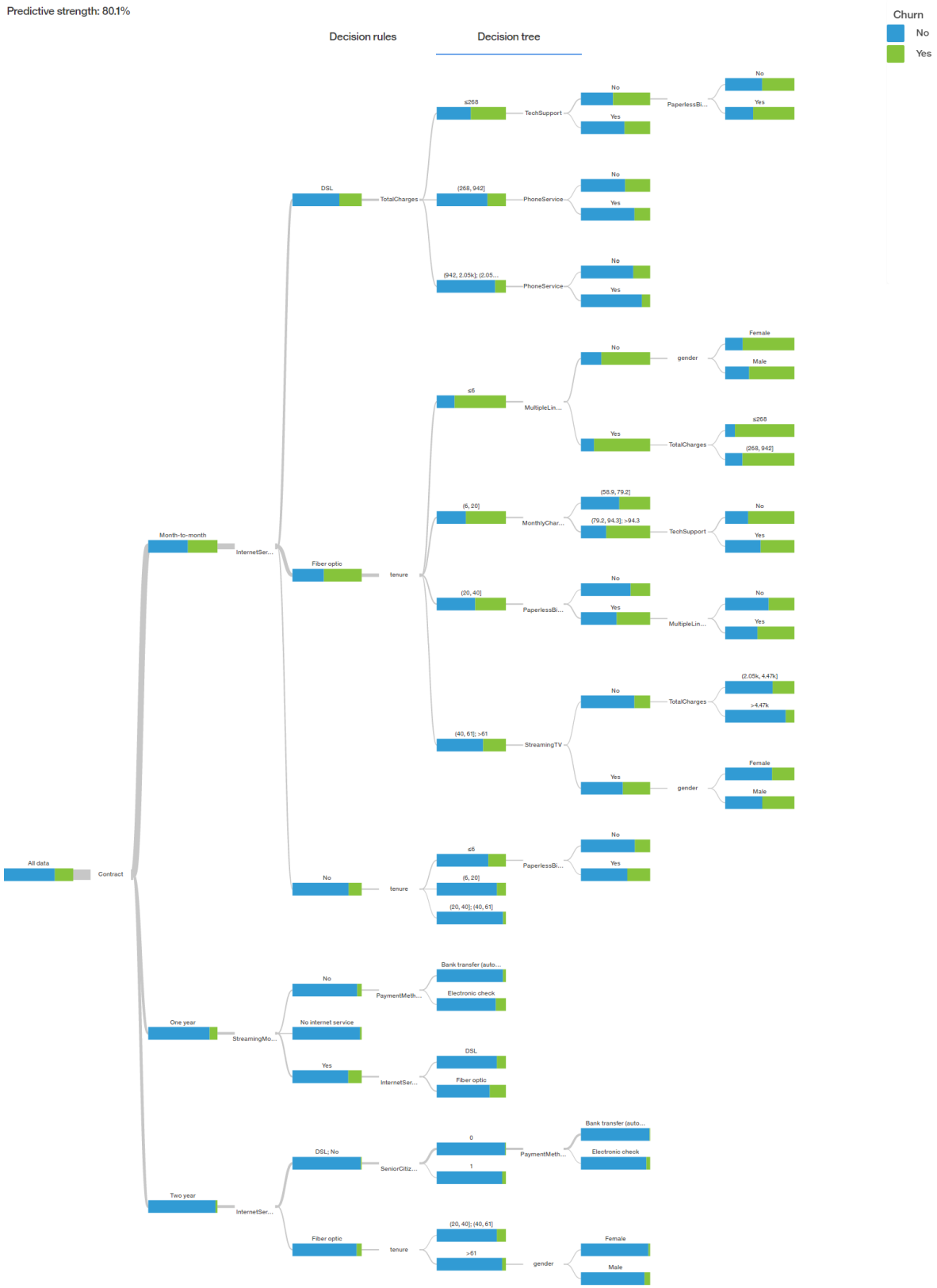

Figure 1. Customer Churn Decision Tree.

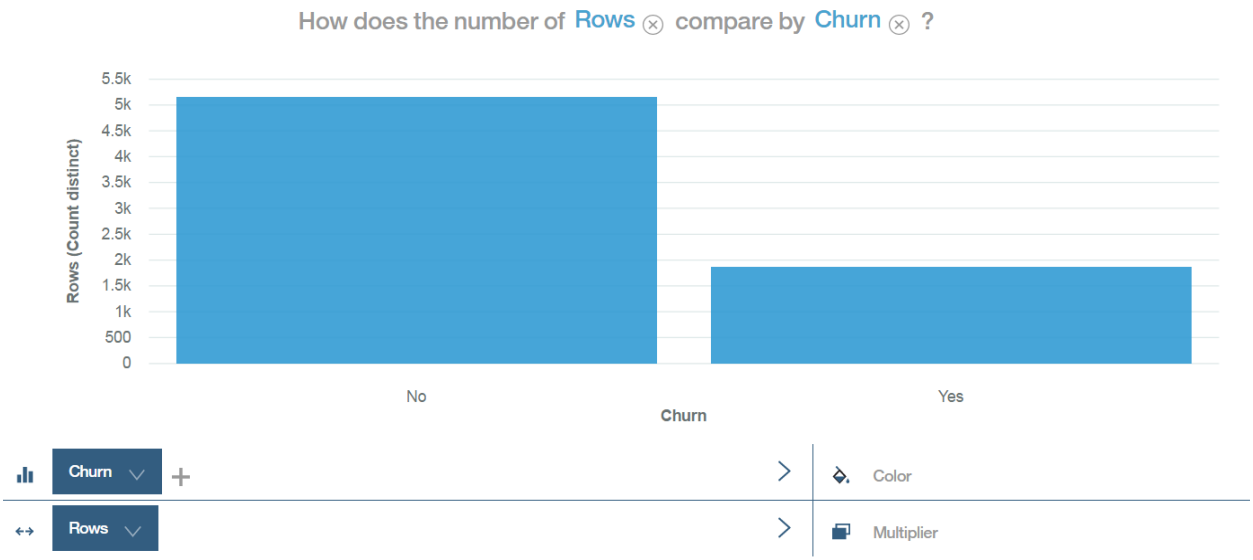

Figure 2. Churn Rate Across All Customers.

Decision rules	Predicted category △▼	Number of records
TotalCharges = ≤268 MultipleLines = Yes tenure = ≤6	Yes | 86%	120
TotalCharges = (268, 942] MultipleLines = Yes tenure = ≤6	Yes | 75%	96
gender = Female MultipleLines = No tenure = ≤6	Yes | 75%	221
TechSupport = No MonthlyCharges = (79.2, 94.3]; >94.3 tenure = (6, 20]	Yes | 67%	340
gender = Male MultipleLines = No tenure = ≤6	Yes | 65%	182

Figure 3. Top Decision Rules for Customers with Highest Churn.

Figure 4. Refined Customer Churn Decision Tree.

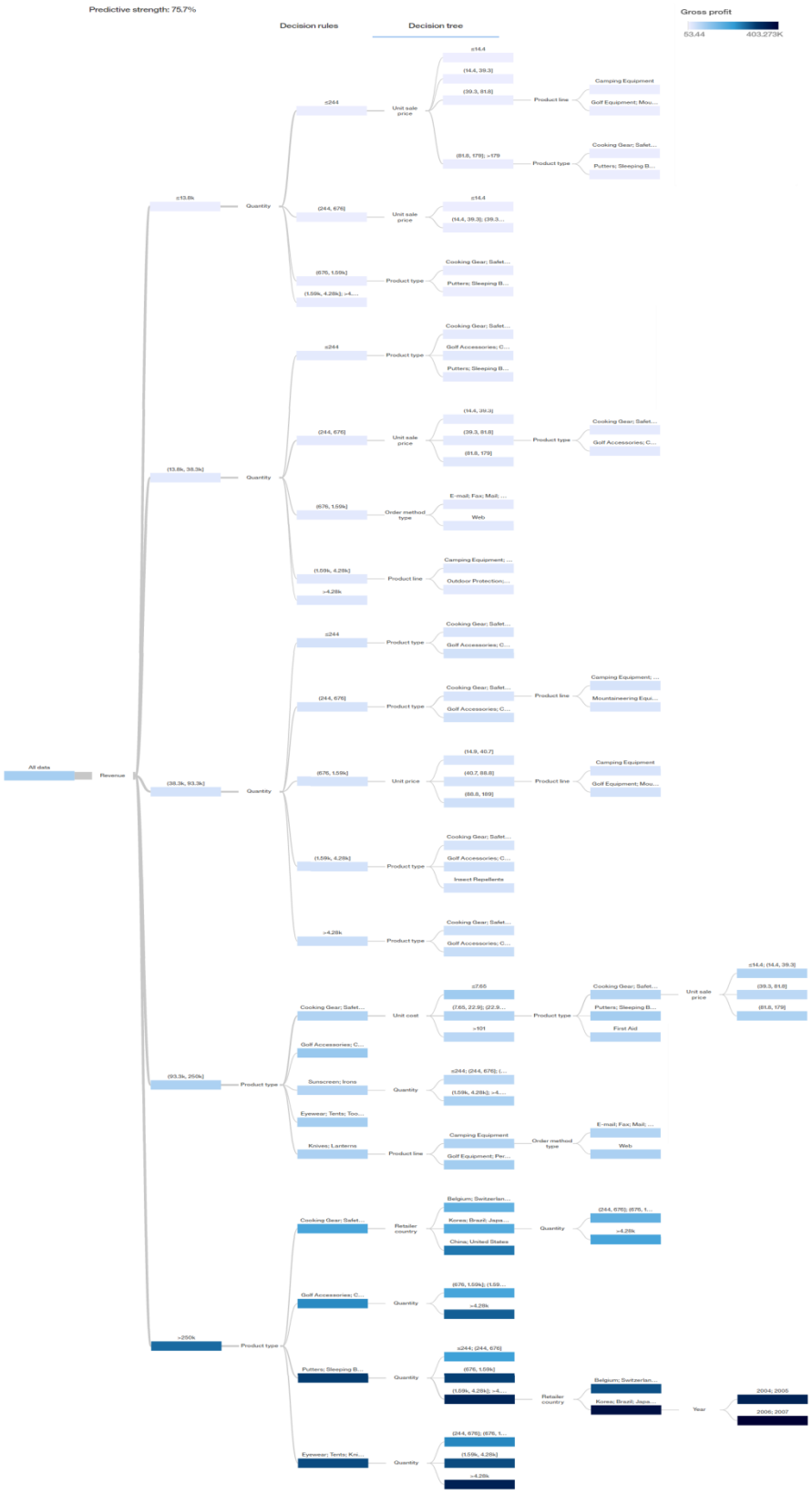

Figure 5. Gross Profit Decision Tree.

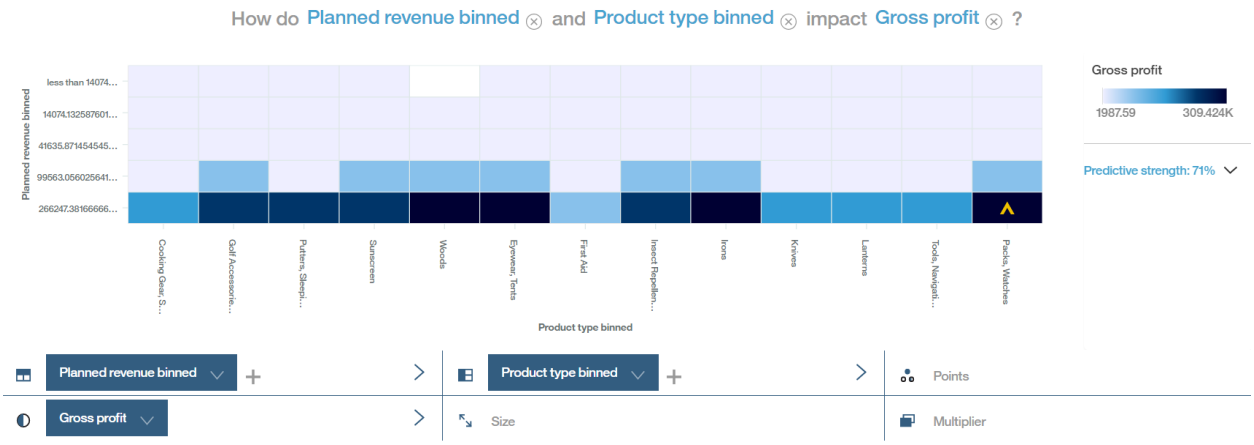

Figure 6. Heatmap Showing Impact of Revenue and Product Type on Gross Profit.

Decision rules	Predicted value △▼	Number of records
Year = 2006; 2007 **Retailer country** = Korea; Brazil; Japan; Canada; France; Singapore; China; United Kingdom; Finland; United States	403273.31	338
Quantity = >4.28k **Product type** = Eyewear; Tents; Knives **Revenue** = >250k	341103.28	437
Year = 2004; 2005 **Retailer country** = Korea; Brazil; Japan; Canada; France; Singapore; China; United Kingdom; Finland; United States	339369.67	291
Quantity = (676, 1.59k] **Product type** = Putters; Sleeping Bags; Packs; Watches **Revenue** = >250k	292097.93	295
Quantity = (1.59k, 4.28k] **Product type** = Eyewear; Tents; Knives **Revenue** = >250k	280846.10	386

Figure 7. Top Decision Rules for Highest Gross Profit.

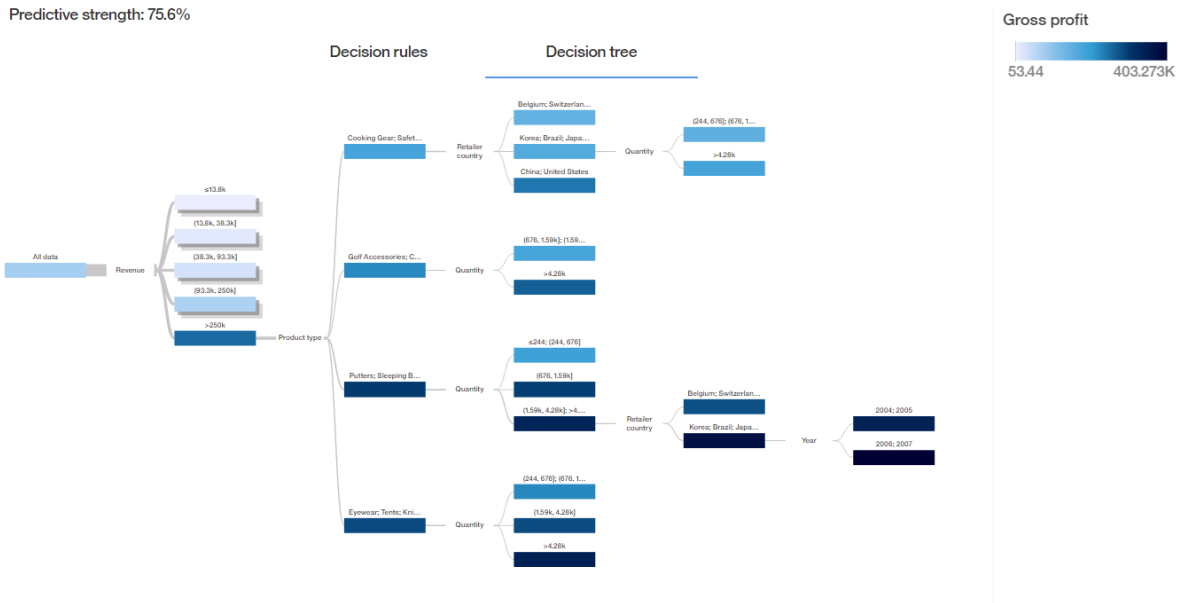

Figure 8. Refined Gross Profit Decision Tree.