Data Management Systems

Course Number: 610

Course Term: Summer, 2020

Assignment number: 3

Project title: Decision Tree Model Development using Cognos Analytics

Contact Information:

Melissa Hunfalvay

Email: [melissa@righteye.com](mailto:melissa@righteye.com)

Phone: 240-357-0026

Instructors Names:

Dr. Moretto

Dr. Goldberg

Introduction

The data set chosen for this assignment is the Billionaires data. This data set represents a multi-decade database of the super-rich. This dataset was gathered from Forbes World’s Billionaires list from 1996-2014 and Peterson Institute for International Economics have added more information about each person on the list. An example of variables added by the Peterson Institute includes whether the billionaires were self-made or inherited their wealth.

The original data set, prior to any cleaning contains 2,614 number of cases or rows of data. There are 22 columns (variables) of data (see Figure 1). Variable include strings, floats, Boolean and integers.

The dataset provides the following variables on the individual that may be used to develop predictive models; age (before cleaning -1 to 98, after cleaning 12 to 98); gender (9.5% female, 90.5% male); the person’s name; their wealth rank; whether they inherited the wealth or not; the region where they live, if they were a founder or were political. Variables also include the wealth type, worth and year the data was reported. Predictive variables related to the citizenship of the individual was included with the largest volume of billionaires from the United States (n = 903) and then Russia (n = 119; see Figure 2).

Details on the company were also reported and can be used as input (predictive) these include company name, date the company was founded and the country code, region, category, sector, type and year the data was collected were all input variables related to the company. For instance, in the variable category the following sectors are reported; finance sectors (30.6%), non-traded sectors (22.8%), traded sector (21.6%), new sector (12.2%), resource sector (9.4%).

Given that predictive models are better developed with larger data sets that have many cases and possible inputs from which to select, only minimal and obvious data cleaning was conducted. The following data cleaning strategies were employed. GDP was removed as 64% of the data was missing. Where age was reported as -1 or 0, it was replaced with the average age of 63 years. Country code was removed as it was redundant with citizenship.

After cleaning was complete there remained 2,614 number of cases or rows of data. There are 20 columns (variables) of unique values with an average of $3.53 worth in billions. Therefore, there is a significant amount of unique data to make the dataset robust for exploration of predictive modeling (see Figure 3).

The intent of the analysis is to determine what input values or factors predict the target variable of worth in billions. Predictive modeling is used on past data, known inputs, to determine future likelihoods of the output variable occurring again. In the case of this dataset, the intent is to use the predictive input variables to determine, via a predictive decision tree model, the likelihood of worth in billions. Perhaps knowing this information with a high degree of accuracy will lead to others, who are looking to become billionaires, to engage in certain similar input variables or activities to increase their future likelihood of becoming a billionaire.

Interestingly at the time of this data compilation Bill Gates was reported as the richest man in the world, worth 76 billion dollars in 2014. Today his net worth is 98 billion dollars and, according to Forbes he is now the second richest man in the world behind Jeff Bezos in 2020 (<https://www.forbes.com/billionaires/>).

Predictive Model Development

Each model was created to predict the target variable, worth in billions. The goal in developing more than one model was to improve upon the predictive strength from model one to model four.

Predictive model one was created by Cognos Analytics after initial data cleaning was conducted. To predict worth in billions the input variables (total of 15) were year, region, gender, citizenship, age, founded, industry, sector, name, rank, wealth type, company type, relationship, was political and category. Results revealed a model that predicted worth in billions using a combination of drivers; rank, year, citizenship, and industry resulting in 68% predictive strength (see Figure 3).

The spiral diagram (Figure 4) shows each driver and combination of drivers that affect the worth in billions. Each driver is represented as a pin or dot and the closer the pin to the center target the higher the predictive strength. Pins are only shown if their predictive strength is 10% or greater.

Decision trees are the most popular predictive analytics tools. They provide rule-based output and are easy to build using computer programming. The decision tree (see Figure 5) diagram is explained by separating the data into subgroups and repeating this process until the tree is complete. The first tree predictor is selected as the top one-way driver. In model one, this is the input variable of rank. Rank is then split into tree nodes showing a rank of less than 174 as the most predictive separation of data (demonstrated by a darker shade of blue, with an average worth in billions of 8.72, standard deviation of 7.8 and total predictive weight in this model of 21.84%). Following this branch of the tree there is a second variable (year) that is then used to split the data further into three years (1996, 2001 and 2014). The year 2014 has the highest predictive value (in dark blue, 7%). The year 2014 was then further split into citizenship with the highest predictors being from Saudi Arabia, United States, Hong Kong, Mexico, Spain, Sweden, Australia and Other, with a predictive strength of 3%. Finally, the nodes end with a final data split at industry with retail, restaurant, consumer, real estate, construction, diversified financial, technology-computer being most predictive within this split, contributing to 2% predictive strength of the overall model.

The sunburst diagram is a different way to visualize of the results of the decision tree (see Figure 6). This diagram represents the target in the middle (worth in billions) with the splits in the data beginning with the first split closest to the target (rank) and subsequent splits moving further away from the target (e.g. year, citizenship). The shade of blue represents the level of strength for each variable.

The rules (see Figure 7) explain details of the decision tree splits with the predictive values on the left side starting with the highest predictors (21.82% for rank of less than 174).The number of records associated with that node is 48 (2%) shown on the far right side of each rule. The rule at the top shows the highest predictive value and then descend to lower values. For instance, in model 1, rule 1 had a predictive value of 21.82%, and included a driver combination of rank, year, citizenship and industry. Rule 2 had the same combination of drivers but different industry related drivers.

In summary, model one shows an overall predictive strength of 68% when combining rank, year, citizenship, and industry. 70% or greater is a benchmark for predictive modeling (Komogortsev & Karpov, 2019). Hence, the model could be improved by further exploration of input variables and combination drivers.

To improve the predictive strength of model one a second decision tree model was created. According to scholars at Peterson Institute for International Economics, approximately half the European billionaires and one-third of the U.S. billionaires got a significant financial boost from family (Anand & Segal, 2017). Therefore, wealth type was filtered to exclude founders, self-made finance and executive. This resulted in an increase in the predictive strength of the model to 75% (see Figure 8).

A third decision tree model was created by including all the inputs in models one, two and by filtering on region and excluding north America, Europe, and Latin America. This resulted in an increase in the predictive strength of the model to 78% (see Figure 9).

A fourth decision tree model was created by including all the inputs in models one, two and three while filtering on year and excluding 2014. This resulted in an increase in the predictive strength of the model to 80% (see Figure 10).

Predictive Model Discussion

Although the predictive strength increased through the model development an interesting counter-effect occurred. These seemingly small changes in inputs caused large changes in the structure of the decision tree (see Figure 11 compared to Figure 5; see Figure 12 compared to Figure 6). Furthermore, such changes would likely result in instability of the model (<https://medium.com/@dhiraj8899/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a>). The instability that could occur in model four that would lead to an inaccurate result is if an individual did not conform to the rank and age drivers of the model yet, were still billionaires. Therefore, although decision tree modeling is popular, we must be careful not to simplify the model so much that is does not account for variation in predicting the worth of billionaires.

The most robust model is model number three, as it has a high predictive score (78%) with a variety of drivers. Model three has the highest predictive value of rank <148 at 6.55. It is a sole driver and top rule. The second rule includes rank between 148 and 311 as well as citizenship, a combination of two drivers with a predictive value of 4.40%.

Some broad similarities were found between models. First the top predictive rule was consistently rank. This makes sense as worth in millions is directly related to your rank on a scale at the time the data was collected. Year and citizenship also contributed to the top five predictive variables for each model. When combining these drivers together they added predictive strength to worth in billions. Inputs that consistently did not contribute to models included gender, age, name, and was political.

More variables including some of the top traits of millionaires (<https://www.investopedia.com/financial-edge/0609/millionaire-traits.aspx#:~:text=Millionaires%20have%20more%20in%20common,skills%20in%20your%20own%20life.>) could be added to the dataset to allow for more robust model development. Diversifying the data set from many billionaire who received money due to family ties compared to millionaires and billionaires who built their wealth would be both more insightful and interesting to organizations as there could be interesting factors they have control over rather than simply luck of an inheritance. The U.S. Trust report data shows that 77 percent of people with at least $3 million in investable assets came from middle class or lower backgrounds while 19 percent surveyed reported growing up poor (<https://www.cnbc.com/2016/05/25/10-traits-rich-people-have-in-common.html>).

Applications to an Organizations Use

The application of this data set to my organization is minimally useful as the results are outputs that are, for the most part, out of ones’ control e.g. inheritance. When viewing the data process from a wider perspective however, the rule-based approach of decision tree modeling could be a very helpful analytical tool to an organization (especially RightEye) as it does not require the normalization of data. Often eye tracking data is skewed, so this would be an advantage of the decision tree analytical process. Another advantage of such rule-based approaches is that missing values does not affect the process of building the decision tree to any considerable extent. Often, data has missing values and therefore this approach would be able to maintain inputs even with less than 100% data present. One of the greatest challenges in adoption of analytics is being able to easily explain results throughout the organization. Decision tree models are easy to explain, they are intuitive.

I found a very interesting comparison between the decision tree analytical model and some rule-based models I have created at RightEye. Instead of using analytics to create models I used scientific logic and rules. I did this as the eye tracking data did not yet exist, however, the clinical and scientific “know-how” was available. Therefore, variables were included in models based on science and splits were determine based on factors such as normality and age. I really appreciate learning this model as now we do have eye tracking data and I may go back and implement this analytical decision tree models for comparative purposes. May the best model win!

A disadvantage of the decision tree modeling analytics is that it would be very time consuming, complex, and expensive to do this type of analysis without a tool like Cognos Analytics. As a small business such a tool is very expensive hence there is a tradeoff between cost of the tool and cost to find another analytical process or tool.

References

Anand, S. & Segal, P. (2017). Who are the global top 1%? *World Development, 95,* p.111-126. [https://doi.org/10.1016/j.worlddev.2017.02.001](https://doi-org.jproxy.lib.ecu.edu/10.1016/j.worlddev.2017.02.001)

Dhiraj, K. (2019). *Top 5 Advantages and Disadvantages of Decision Tree Algorithm.* Retrieved from: <https://medium.com/@dhiraj8899/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a>

Dolan, K.A. (2020). *Forbes World’s Billionaire List.* Retrieved from <https://www.forbes.com/billionaires/>

Komogortsev, O.V. & Karpov, A. (2019). Automated classification and scoring of smooth pursuit eye movements 382 in the presence of fixations and saccades. *Behavior Research Methods*, 45(1). https://doi.org/ [10.3758/s13428-012-0234-9](https://dx.doi.org/10.3758/s13428-012-0234-9)

Little, K. (2016). 10 Traits Rich People Have in Common. Retrieved from: <https://www.cnbc.com/2016/05/25/10-traits-rich-people-have-in-common.html>

Powers, S. (2019). *6 Millionaire Traits That You Can Adopt.* Retrieved from: <https://medium.com/@dhiraj8899/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a>

Appendix A

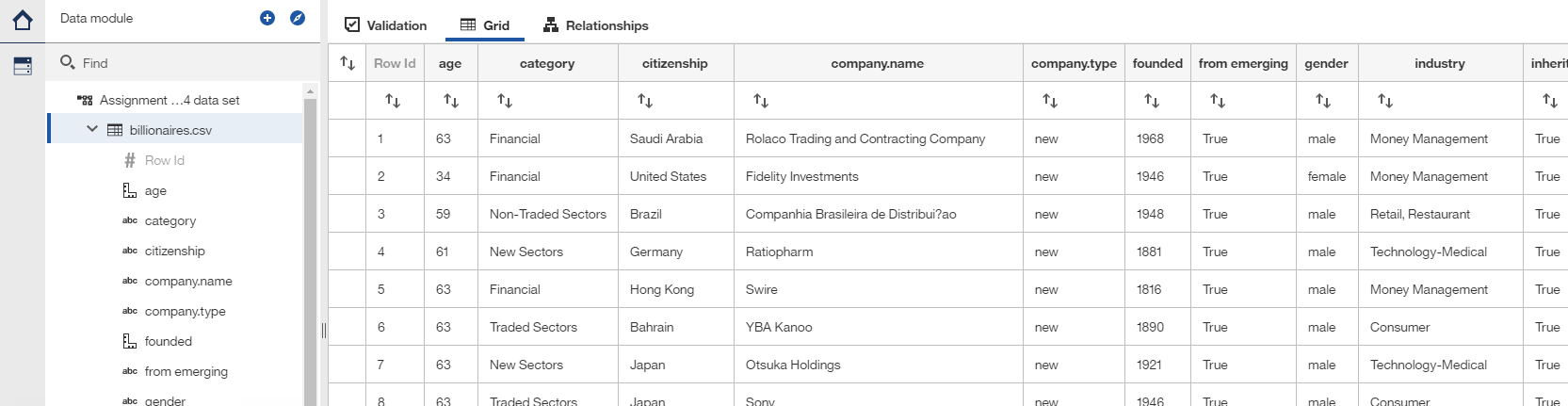


Figure 1: Billionaires data set

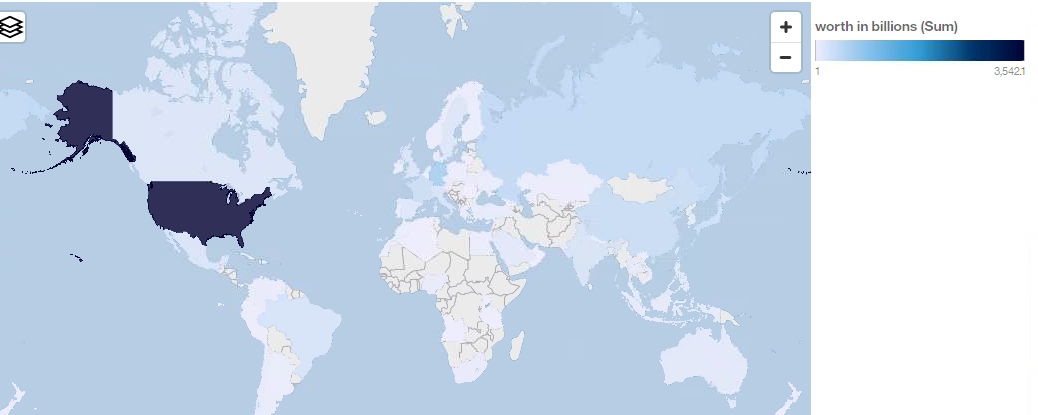


Figure 2: Worth in billions by citizenship

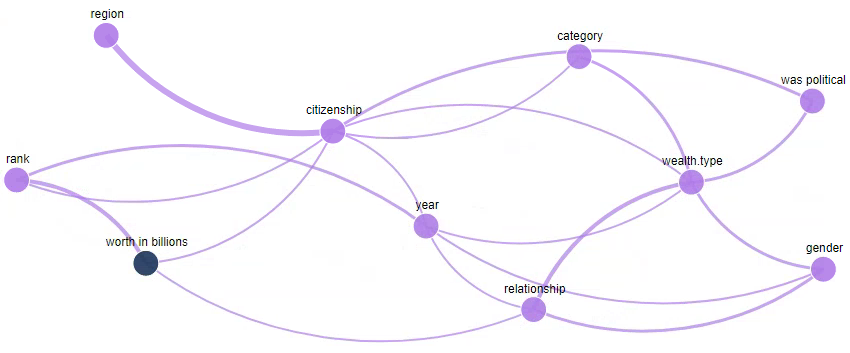


Figure 3: Relationship map after data cleaning

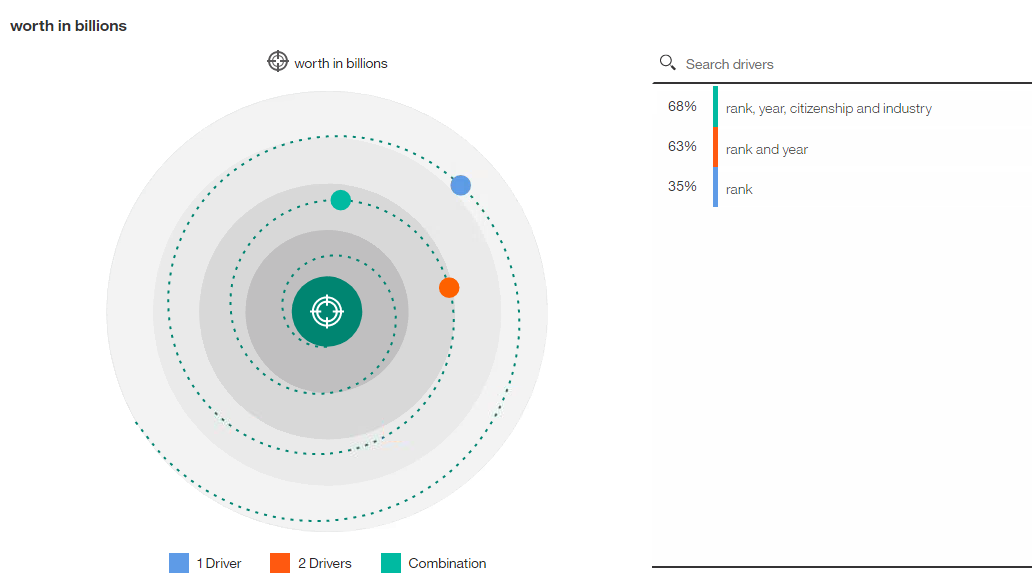


Figure 4: Model one displaying a combination of rank, year, citizenship, and industry predicting worth in billions with a predictive strength of 68%

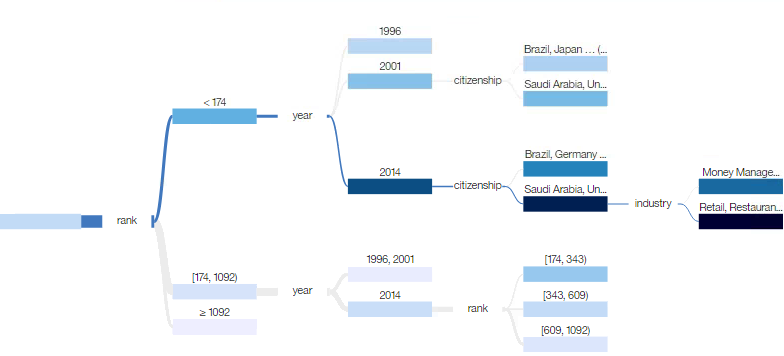
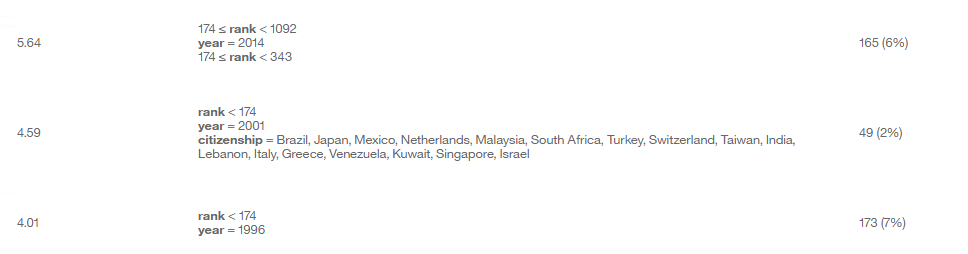


Figure 5: Model one decision tree



Figure 6: Model one sunburst diagram





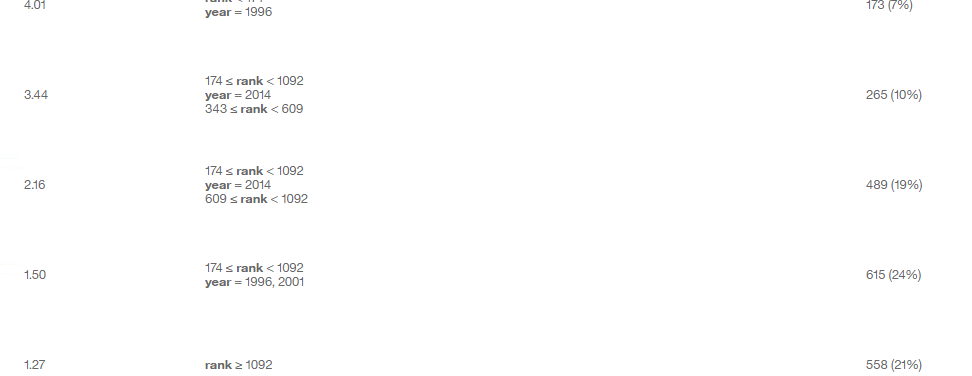


Figure 7: Model one rules

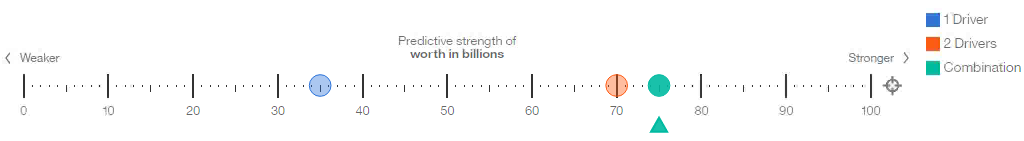


Figure 8: Model two excluding founders, non-finance, self-made finance, and executives

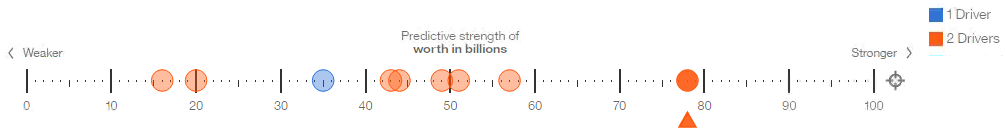


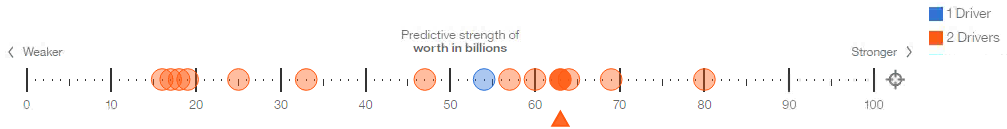
Figure 9: Model three excluding regions of North America, Europe, and Latin America

Figure 10: Model four filtering on year and excluding 2014

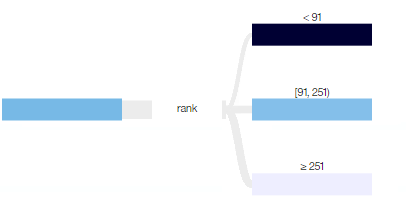


Figure 11: Model four decision tree diagram

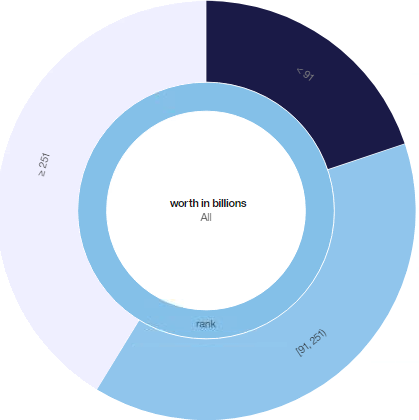


Figure 12: Model four sunburst diagram