Data 610: Decision Management Systems Fall 2022

Assignment # 3 – Decision Tree model development using Cognos Analytics

Simon A Hochmuth

Email: shochmuth3@student.umgc.edu

Professor: George Cross

Introduction

The billionaire CSV file has data surrounding billionaires across the world over multiple decades with the goal of understanding trends among the super-rich. The foundation of the data was from the Forbes World Billionaires list between 1996 to 2014, and researchers from Peterson Institute for International Economics have added multiple variables on top of the original Forbes data. The dataset contains thirteen strings, three booleans, four integers, and two float data types (Whitcomb. 2016). The total number of records within the dataset is 2614, with twenty-two columns. The variables within the dataset include information on a billionaire's companies, industry, location, age, rank, demographics, and the year the data was collected. The goal of this dataset is to predict the target variable, called worth in billions.

Data Preparation

Figure 1 shows that the columns relationship, sector, name, wealth type, gender, wealth category, and industry are missing values. Each column is missing at most 1.4% of the data. Therefore, the best way to deal with this data will be to filter out the nulls since it will not take away from the trends within the data. For that reason, the final data module used to create the decision trees will include each column with missing data filtered out.

The next phase of preparing the data included exploring the data to find the columns considered irrelevant. For example, the columns inherited from emerging and was founder are insignificant since the columns only contain one unique value. In addition, the columns citizenship and sector were irrelevant because they duplicated country code and industry. After

removing these columns, the final module shown in Figure 2 contains seventeen columns used to create a decision tree that predicts the target variable. The next step will be to use Cognos Analytics to find which of these variables are drivers for the target variable worth in billions and create decision trees to model this predictive model.

Develop and discuss at least two predictive models and the results

Decision tree model on the original dataset

Initially, the models shown in Figures 3 and 4 show the model on the original data with no filters applied. The original data was used to see how Cognos would analyze it to understand changes for the next model. The first thing to note is that no decision tree model will work if the column rank is filtered out. Rank is the primary variable each decision tree requires to create a predictive model and why rank will be in every model. After the rank is analyzed, the original model splits between the column's year, wealth type, and country code. These splits show that each of these variables is crucial in understanding the data set. This model has a 68.4 % predictive strength, according to Cognos. However, the goal in future models will be to remove or filter these three columns to see how the strength changes. The Sunburst diagram in Figure 4 shows that the top five target values range from five to twenty billion, whereas the sunburst diagram is longer when predicting the highest net worth. The extra nodes lengthen specific sections of the diagram, which depict that the model requires more complicated calculations to predict these target ranges. The highest net worth is shown in the darker blue section of figure 4 and leads to the question of what would happen if we filtered the rank or worth in billions to not focus on outliers. Therefore, the first step will be to create a model excluding wealthier regions to see the effect on the overall predictive strength.

Decision tree model excluding North America and Europe

From the model discussed previously, one of the nodes that led to the predictive model was the country code. The question then arose, what would happen if North America and Europe were filtered out since these regions are doing better than the rest of the world? The filter will show if wealthier regions skew the predictive strength of the model. Figures 5 and 6 show the model with these two regions excluded. The new model had a predictive strength of 75.9%, increasing from the previous model. That shows that Cognos is more confident in the data coming from the non-excluded countries. The sunburst diagram in figure 6 shows that the wealthiest people were in 2014, while more nodes were required to predict the individuals in the target range of three to five billion. The diagram showed that the column year is an essential predictive driver, and a future model will focus on how predictive strength changes when the year is filtered. The model also has a top target value of thirteen billion, which is significantly smaller than the model in Figure 3, with a top target value of twenty-two billion. Both models have a top target value less than the list's highest net worth of seventy-six billion. Another model should filter the worth in billions to see how predictive strength changes when removing outliers from the target variable.

Top five rules found in initial model analysis

As previously discussed, Figures 3 and 4 show clear trends in predicting a person's worth in billions. Overall, the trends in the first model were related to the column's rank, year, country code, and wealth type. The first major trend in both models was that rank was a critical predictive driver. The rank went from 1 to 1565, and splitting it into sections helped build predictive models. This makes sense because people with similar ranks likely have a similar net

worth. The following common trend was that the model saw a difference in net worth between the three years 1996, 2001, and 2014 where many predictions were made by pairing year with rank. Using year as a predictive driver makes sense because, due to inflation, the world's money will go up each year, creating wealthier billionaires. After rules for a year are applied, rank is used to further model the data. The last two variables were applied to country code and wealth type. Specifically, when looking at the highest earners with a rank under 180, the model required a more complex rule to predict net worth. All these variables make sense for predictions because the country's location could favor more affluent areas, and certain wealth types can produce more billionaires. Using these trends, the model created the top rules shown in Figure 7.

Discussion of additional models developed

Decision tree model for the year 2014

Figures 3 to 6 show that the variable year was a clear driver in the decision tree model. For that reason, an additional model was made, excluding the years 1996 and 2001, to see if the model has more robust results when looking at a single year. Since the original data originated from the Forbes yearly billionaire list, a hypothesis is that each year has its own set of trends. By filtering to 2014, it is shown in Figure 8 that the predictive strength of the model significantly decreases to 58.5%. The fact that the predictive strength significantly decreases shows that Cognos does not see underlying trends each year. This model also shows a target variable that does not exceed twenty billion dollars, showing that the highest earners are less represented. This leads to the question, will filtering the target variable to remove outliers increase the overall predictive strength of the model?

Decision tree model filtering worth in billions

All previous models have shown that the highest target value was a little under twenty-five billion, which led to the creation of the model in Figure 9. The filter decreased the highest net worth from seventy-six to twenty-five billion. Doing this increased the model's strength to 76.5% without any other filtering done to the data set. This dramatic increase shows the impact of outliers within the target variable on the overall model. The reason it decreased to twenty-five billion was that every model before never had a target value higher than that. Showing that Cognos lost predictive strength the higher the worth in billions went. This model still utilized the variables rank, year, and country code, but it could predict with higher confidence due to the missing outliers. Overall, this model does a great job of predicting the target variable under twenty-five billion.

Outline and discuss a rules-based approach in your organization

Booz Allen Hamilton is a contracting firm that works within the Department of Defense (DOD) that specializes in creating a shared data storage location for pipelines and creating analytical tools. A rules-based decision approach would be a perfect fit for this corporation that is deeply embedded within the government because any small change can scale to an organization that affects every American.

Rules-based approach with maintenance data

In 2021 the DOD spent $718 billion on military activities, where 40% of the budget went into operations and maintenance (PGP Foundation. 2021). Because maintenance is such a large part of the budget, leadership needs to decrease the inefficiencies in the business. In the organization,

some data sets pull from multiple military branches and bases worldwide, which track a broad range of equipment from guns to vehicles. All of these are vital to ensuring the warfighter is mission ready. Since the military is large, plenty of data can be used to increase efficiency within each branch, including its equipment stockpile. This data is naturally classified, so an example of this approach will be made with an organization like the US Postal Service (USPS).

Every American has seen a USPS truck on a mail route but does not realize the USPS had over 211000 carriers delivering mail to eighty-seven million delivery points in 2008, with an average of seven hundred delivery points a day. (Office of the Inspector General, 2008). With all the various locations across America, each truck is experiencing different climates, terrain, mileage, speeds, or drivers. All these variables can go into predicting whether a truck needs preventative maintenance. A simple example is that trucks delivering mail in the Colorado mountains in the winter will have more maintenance than trucks in flat Iowa during the spring. The target variable will be to predict the time needed between maintenance cycles to ensure the fleet does not waste money or destroy equipment. In this example, location plays a significant part in the calculation because it determines climate, terrain, and even the management in charge. For that reason, this is likely the first predictive driver for this model. Next, the variables dealing with terrain specifically would be necessary since each route at a location can be different. By tracking the altitude or road types, a grade can describe a route's terrain, which can decide whether particular trucks are experiencing more stress. Lastly, if a driver is cautious or aggressive, it can determine failure. Since the stress on vehicles would be higher if someone uses more brakes or accelerates at higher rates. These stress levels could be used to see who within a team of drivers is more likely to have unsafe driving practices by checking things like speed, acceleration, or brakes. These three points are part of a simplified decision tree model for

predicting the time between maintenance. Added variables to the USPS model could include miles driven per route, type of road driven on, average oil pressure, average temperature, or even the vehicle brand. All these variables are relevant to the DOD and can be a part of the predictive maintenance model for the military branches. Other DOD specific variables are country, rate of fire for weapons, wartime status, ammunition type, or training frequency. This model would be highly effective in saving maintenance costs, saving millions of taxpayer dollars each year.

References

Whitcomb, R. (2016, May 17). *Billionaires CSV File*. Tools | ct-vt.github.io. https://think.cs.vt.edu/corgis/csv/billionaires/

Office of the Inspector General. (2009, March). City delivery route consolidation | USPS Office of Inspector General. https://www.uspsoig.gov/blog/city-delivery-route-consolidation

PGP Foundation. (2022, June 1). Budget basics: National defense. Peter G. Peterson Foundation. https://www.pgpf.org/budget-basics/budget-explainer-national-defense

Appendix A

```
In [5]:  1  df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2614 entries, 0 to 2613
Data columns (total 22 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   name                      2614 non-null   object
 1   rank                      2614 non-null   int64
 2   year                      2614 non-null   int64
 3   company.founded           2614 non-null   int64
 4   company.name              2576 non-null   object
 5   company.relationship      2568 non-null   object
 6   company.sector            2591 non-null   object
 7   company.type              2578 non-null   object
 8   demographics.age          2614 non-null   int64
 9   demographics.gender       2580 non-null   object
 10  location.citizenship      2614 non-null   object
 11  location.country code     2614 non-null   object
 12  location.gdp              2614 non-null   float64
 13  location.region           2614 non-null   object
 14  wealth.type               2592 non-null   object
 15  wealth.worth in billions  2614 non-null   float64
 16  wealth.how.category       2613 non-null   object
 17  wealth.how.from emerging  2614 non-null   bool
 18  wealth.how.industry       2613 non-null   object
 19  wealth.how.inherited      2614 non-null   object
 20  wealth.how.was founder    2614 non-null   bool
 21  wealth.how.was political  2614 non-null   bool
dtypes: bool(3), float64(2), int64(4), object(13)
memory usage: 395.8+ KB
```

Figure 1: ***Python summary of the billionaires.csv dataset showing datatypes and null values***

age	category	company.name	compa...type	country code	founded	gdp	worth i...illions	gender
-1	Financial	Rolaco Trading and Contracting Company	new	SAU	1968	158000000000	1	male
34	Financial	Fidelity Investments	new	USA	1946	8100000000000	2.5	female
59	Non-Traded Sectors	Companhia Brasileira de Distribui?ao	new	BRA	1948	854000000000	1.2	male
61	New Sectors	Ratiopharm	new	DEU	1881	2500000000000	1	male
-1	Financial	Swire	new	HKG	1816	160000000000	2.2	male
-1	Traded Sectors	YBA Kanoo	new	BHR	1890	6100000000	1	male
-1	New Sectors	Otsuka Holdings	new	JPN	1921	4710000000000	2.2	male
-1	Traded Sectors	Sony	new	JPN	1946	4710000000000	2.3	male
66	Financial	Mori Building	new	JPN	1959	4710000000000	3.9	male
-1	Traded Sectors	Chanel	new	FRA	1909	1610000000000	2	male
12	Financial	von Thurn and Taxis family	new	DEU	1615	2500000000000	1.5	male
63	Resource Related	Penoles	privatization	MEX	1960	397000000000	1.8	male
-1	Non-Traded Sectors	Groupo IUSA	new	MEX	1939	397000000000	2.5	male
-1	Financial	Pulsar International	new	MEX	1981	397000000000	1.4	male

Figure 2: Image of billionaire dataset that includes relevant variables.

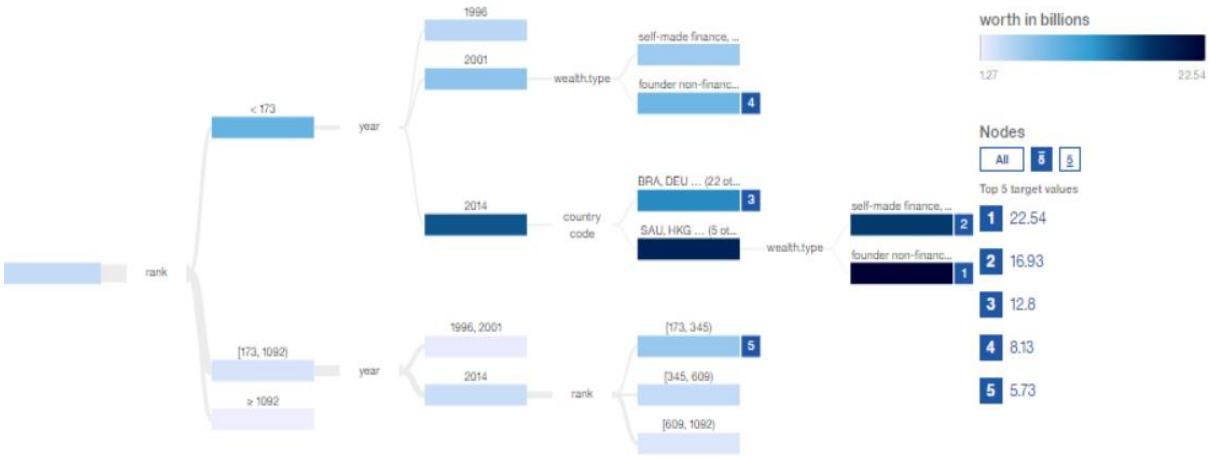

Figure 3: ***Decision tree diagram where the decision points include rank, year, country code, and wealth type. This tree is made without filters on the original dataset and has a 68.3% predictive strength.***

Figure 4: ***Sunburst diagram from the model shown in Figure 3.***

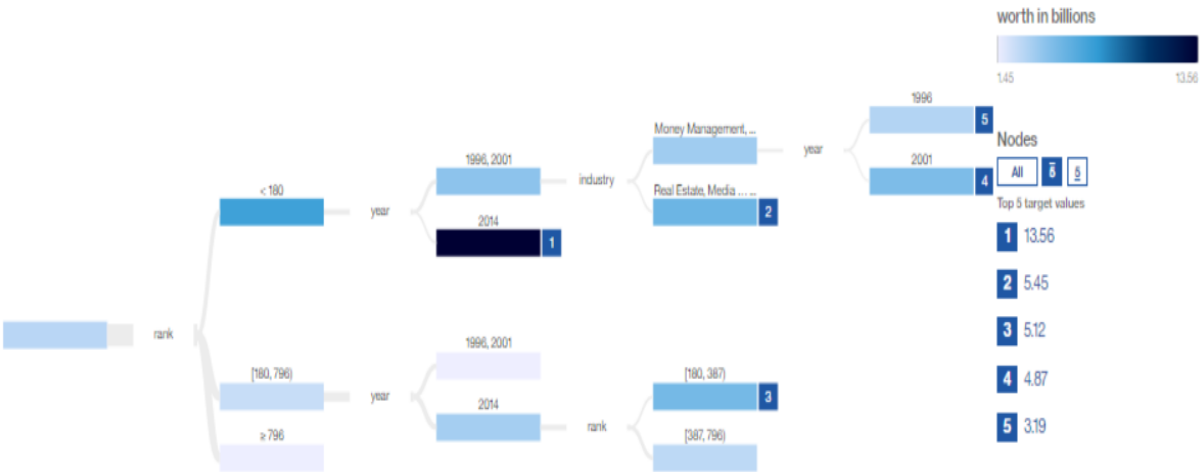

Figure 5: ***Decision tree diagram where the decision points include rank, year, country code, and wealth type. This is a tree made by excluding Europe and North American regions and has a 75.9% predictive strength.***

Figure 6: ***Sunburst Diagram from the model shown in Figure 5.***

ΔPredicted value	Rules	Records
22.54	**rank** < 173 **year** = 2014 **country code** = SAU, HKG, MEX, USA, ESP, SWE, Other **wealth.type** = founder non-finance, privatized and resources	25 (1%)
16.93	**rank** < 173 **year** = 2014 **country code** = SAU, HKG, MEX, USA, ESP, SWE, Other **wealth.type** = self-made finance, inherited, executive	42 (2%)
12.80	**rank** < 173 **year** = 2014 **country code** = BRA, DEU, JPN, FRA, MYS, PHL, THA, CHE, COL, Taiwan, CAN, KOR, GBR, IND, ITA, IRL, SGP, RUS, CHN, DEN, CZE, AUT, UKR, CYP	87 (4%)
8.13	**rank** < 173 **year** = 2001 **wealth.type** = founder non-finance, privatized and resources	50 (2%)
5.73	173 ≤ **rank** < 1092 **year** = 2014 173 ≤ **rank** < 345	147 (6%)
5.46	**rank** < 173 **year** = 2001 **wealth.type** = self-made finance, inherited, executive	109 (5%)

Figure 7: ***Top 5 rules for models shown in Figures 3 and 4.***

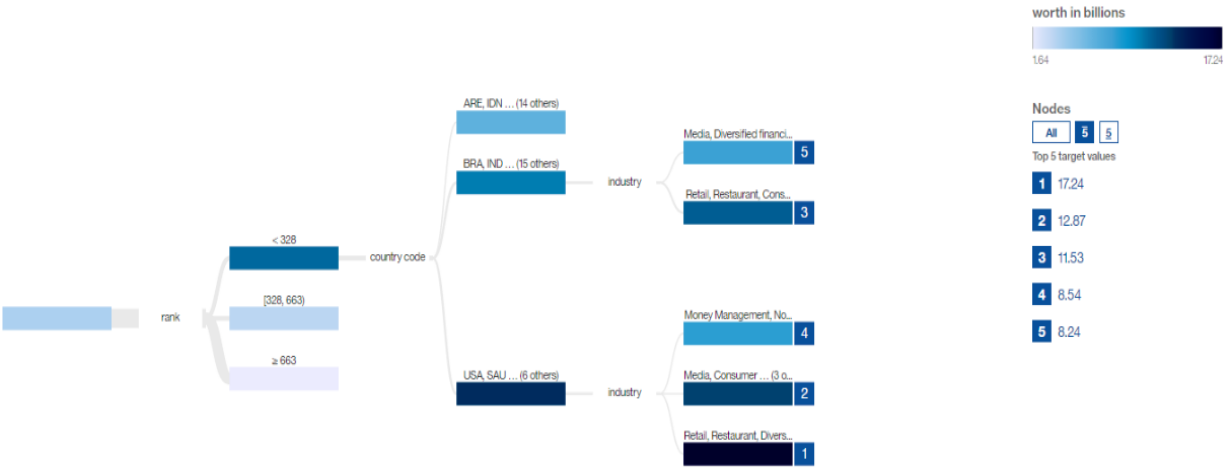

Figure 8: ***Decision tree diagram where the decision points include the column's rank, country code, and industry. This is a tree made filtering in 2014 and has a 58.5% predictive strength.***

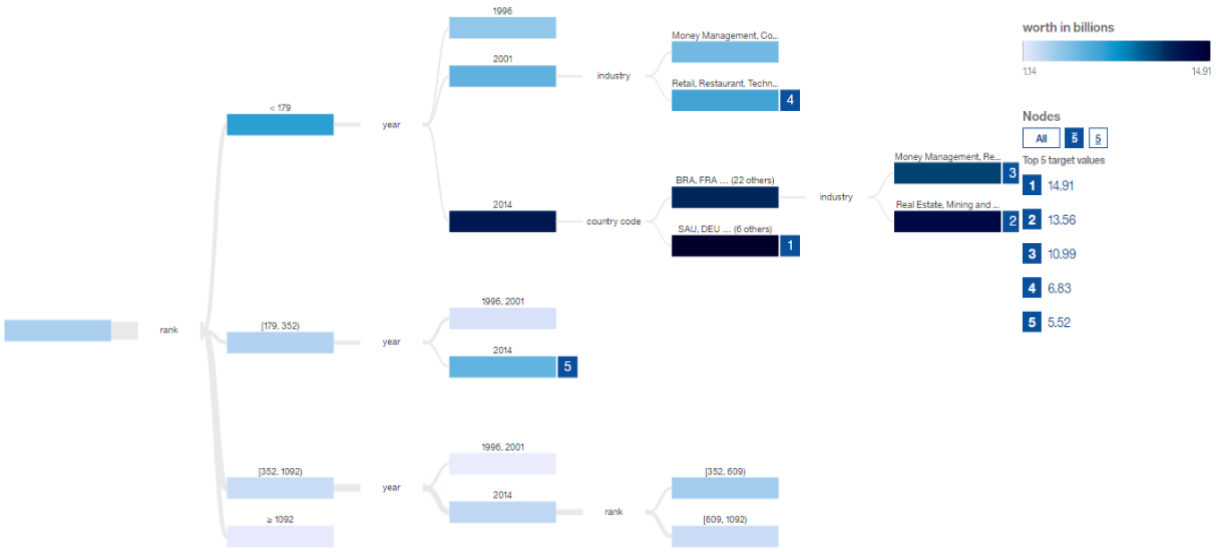

Figure 9: ***Decision tree diagram where the decision points include rank, year, country code, and industry. This is a tree made filtering the column worth in billions to under 25 billion. This model has a 76.5% predictive strength.***