Suffolk University

Boston, MA 

Final Project

ISOM 330 AE: Applied Stats and Predictive Analytics

April 26, 2023

\*The story and the purpose of this model is to learn about billionaires around the world. We want to learn how the relationships between their personal background affect or relate to each other. Furthermore, exploring if there are trends considering education level, background, and family status.

Dataset: Forbes Billionaires - Forbes' Dataset comprises 2750 people whose net wealth equals to or exceeds one billion US Dollars.

Variables:

1. Name
2. Net Worth: expressed in USD
3. Country of Origin
4. Source: main source from which the wealth is originated
5. Rank: position in terms of wealth
6. Age
7. Residence: city where the billionaire lives
8. Citizenship
9. Marital Status
10. Children: number of children the billionaire has
11. Education: level of education they achieved
12. Self-Made (True/False) origin of wealth
13. Geometry: coordinate of current residence

* First, we used the readr library to import our Forbes Billionaire dataset and name its Billionaire.
* From the summary of the dataset, we can notice that the following variables have missing values: Age has 125 NA’s, Children has 1203 NA’s, Self-made has 18NA’s
* We use kNN from VIM library to impute missing values with k=1
* After, we remove the zero and nearzero variances using the caret library
* We create dummy variables for Country, Education, Status:
  + Country: If the country of origin is United States the value is 1, else is 0

The dataset includes many countries around the world that billionaires are from. The majority of billionaires are located in the US. So we grouped it as “in the US” and “outside of US”

* + Education: If it contains the word “Bachelor” the value is 1, else is 0

The dataset includes many Education level” PhD, dropout, high school, etc. However, the majority had a bachelor’s degree so that is why we grouped it as “Bachelor” and not

* + Status: If it contains the word “Married or Remarried” the value is 1, else is 0

The data set includes many status such as divorce, dating, married, etc. but the majority are in a marriage so we grouped it as “Married” and not.

* We delete the Country, Education, Status variables.
* Before we make any changes to the data, we run a linear regression with all the variables to check for significant variables.
* Based on the summary we can conclude that Children and Education are significant to net worth and they are both positive which means the higher the number of children a billionaire has, the higher their net worth. And billionaires with a bachelor's have higher net worth than the ones without.
* Running a second linear regression only with the significant variables, a third one with variable transformation (taking the log of the response variable) and a fourth one with variable transformation (taking the sqrt of the response variable). We noticed that based on the adjusted R squared our third linear regression is a slightly better model than the rest.

1. Linear Regression Model

\*We deleted the text variables: name, source residence, citizenship

* 1. lm.Billionaire: running on all predictors

Based on the summary we see that Children and Education are significant to Networth. They both have positive coefficients, which mean that the higher the number of children the billionaire has, the higher their networth. Also, the billionaires with a bachelor's degree have higher networth than the ones who do not have it.

Every 1 unit increases in children, there is 0.00873 unit increase in networth. Every unit increase in education, there is 8.85e-06 unit increase in networth.

\*After the first linear regression, we removed all non-significant variables which left two significant variables which are Children and Education

* 1. lm.Billionaire2: running on the significant variables

Both Children and Education are still significant to Networth

Every 1 unit increases in children, there is 0.00229 unit increase in Networth. Every unit increase in education, there is 2.15e-08 unit increase in Networth.

* 1. lm.Billionaire3 and lm.Billionaire4: running transformation on Log and Sqrt

lm.Billionaire2, lm.Billionaire3, lm.Billionaire4 with adjusted R squared respectively: 0.01439, 0.02944, 0.02339. lm.Billionaire3 with the Log transformation has the highest adjusted R square value which means this model is a better fit than the rest.

* 1. Interpret diagnostic plots:
  + Residuals and Fitted: The residuals are not equally spread around the horizontal lines, suggesting non-linear relationships were not explained by the model.
  + Normal Q-Q: The residuals deviate severely at the 2/3 of the plot so this plot is not good. They are not lined well on the straight dashed line.
  + Scale-Location: The residuals do not appear randomly or equally along the horizontal line. They focus densely along the x-axis as it passes from 0-8. This plot is not good either.

1. Classification Model
2. glm.Billionaire: running for the relationship between education and Networth. From the summary we see that Networth is significant and positive. When the NetWorth increases so does the probability that the billionaires have a bachelor's degree.
3. glm.Billionaire2: running the relationship between Country and Status. From the summary we notice that Status is significant and positive which means, billionaires who are married have a higher probability that they are from USA than not married people.
4. Diving The Dataset

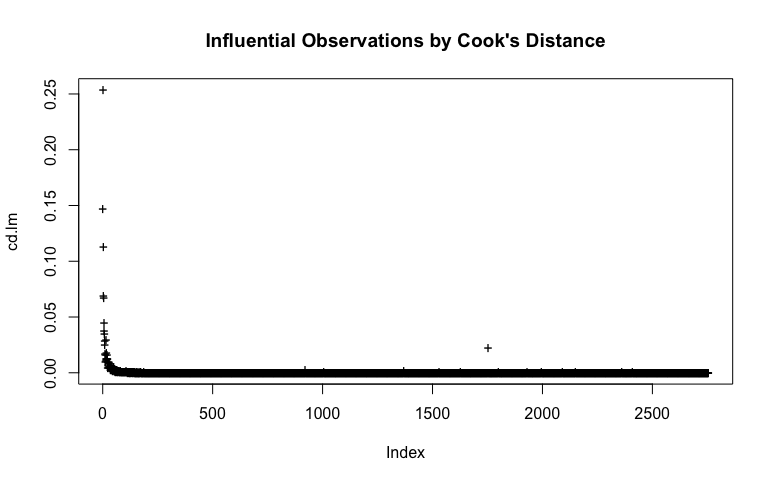
The Dataset was divided into 1755 to 1000. Where 1755 data record was used to make training dataset as we want the predict models to be accurate. Whereas 1000 records was used to make the testing dataset as it can be used to have a comparison with the training dataset

1. KNN Model

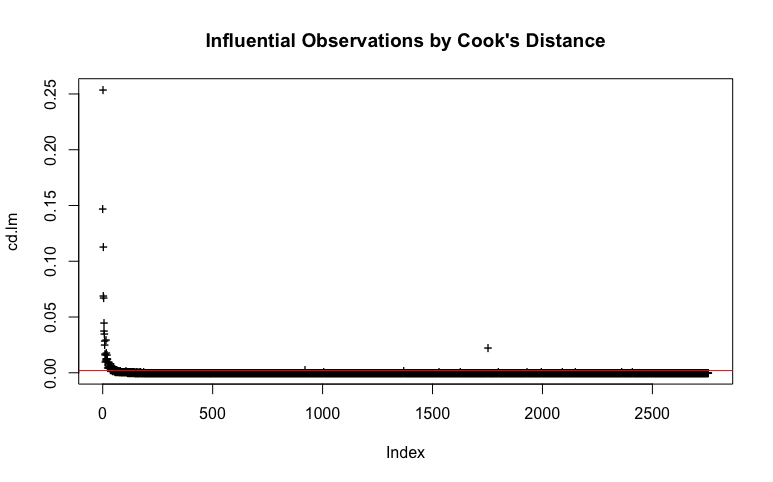
The KNN model is used to impute the missing values with the K nearest neighbor. Three kNN models were developed in order to find the best and most efficient model. Where K=1, K=3 and k=5 was used. While comparing the error rate it was concluded that the K=1 has the least error rate as compared to the other kNN models which was developed.

1. Graphs

a)Cook Distance



b)Cook Distance with Red Abline



c)Correlation Graph

Chart, bubble chart

Description automatically generated

d)Diagnostic Plot

