Teranalytics Test

The voting problem

objective:

Predict if a citizen will vote based on his age and peers, and analyzing the data.

César Vega

Contents

Executive summary

By analyzing the social and age groups of voters, we built a model that could predict the if a citizen would vote with 87% accuracy. Insights on the data and models are also found in this document.

Objective

Predict if a citizen will vote based on his age and peers, and analyzing the data.

Task 1- Data exploration

Q1 – Of the voters in the **groups.csv** file, how many of them are not in the voting record file –

- 73%

Q2- Of the voters in the **groups.csv** file for which we have a voting record what percentage of them voted in the year 2006?

- 26%

Q3 - Are the 5 groups exclusive of each other? How many voters for which we have voter records are listed as in multiple groups?

- The groups are exclusive to each other
- There were 11 voters listed as duplicates

Q4 - Of the remaining voters, what percentage of them are in each of the following age groups:

- 18-30: 8%
- 31-40: 13%
- 41-50: 28%
- 51-65: 36%
- 65-all: 15%

Q5 - For each age group, what percentage of people voted in 2006? Which age group had the largest percentage of people vote?

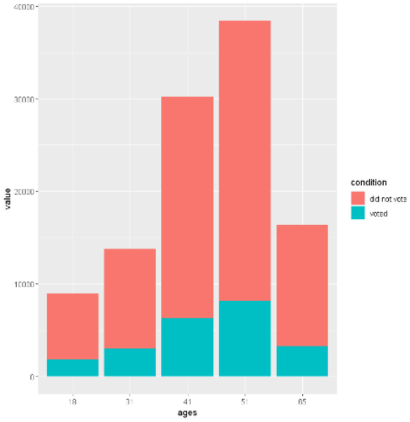
- 18-30: 2%
- 31-40: 4%
- 41-50: 7%
- 51-65: 10%
- 65-all: 4%

Q6 - For each treatment group, what percentage of people voted in 2006? Which treatment group had the largest percentage of people vote?

- civic: 6%
- Hawthorne: 5%
- self: 6%
- neighbors: 6%
- control: 4%

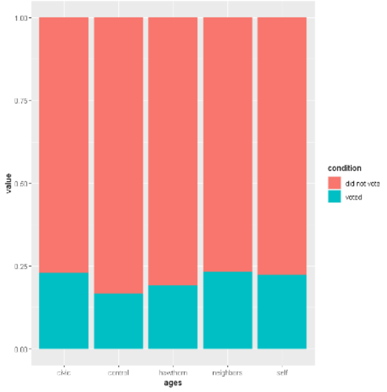
Task 2 – Data visualization

Q1. Stacked bar chart – voter breakdown



-

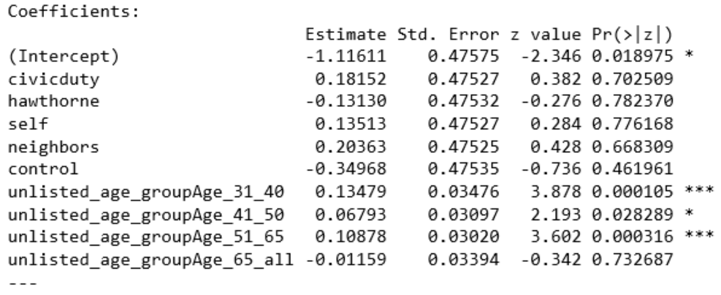
Q2. Stacked bar chart – group breakdown



-

Task 3 – Modeling with logistic regression

Part 1 – Our network coefficients:



```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)    
(Intercept)                   -1.11611    0.47575  -2.346 0.018975 *  
civicduty                      0.18152    0.47527   0.382 0.702509    
hawthorne                     -0.13130    0.47532  -0.276 0.782370    
self                           0.13513    0.47527   0.284 0.776168    
neighbors                      0.20363    0.47525   0.428 0.668309    
control                       -0.34968    0.47535  -0.736 0.461961    
unlisted_age_groupAge_31_40    0.13479    0.03476   3.878 0.000105 ***
unlisted_age_groupAge_41_50    0.06793    0.03097   2.193 0.028289 *  
unlisted_age_groupAge_51_65    0.10878    0.03020   3.602 0.000316 ***
unlisted_age_groupAge_65_all  -0.01159    0.03394  -0.342 0.732687    
---
```

-

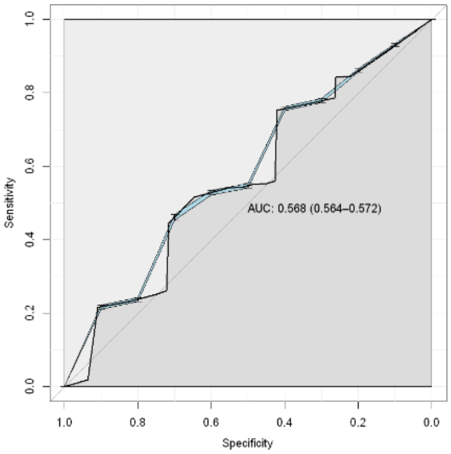
- In the output above, we see the deviance residuals, which are a measure of model fit. This part of output shows the distribution of the deviance residuals for individual cases used in the model. – The coefficient with most significance in this regression model is the p-value Pr(>z), which indicates evidence against or for a null hypothesis.

- Next it shows the coefficients, standard errors, the z-statistic, and the associated p-values, here we can see that the group values are not statistically significand, but every age value is (except for age 65+)

Part 2- Using a threshold of 0.5, what is the accuracy of the logistic regression model.

- Here I used two ways of calculating accuracy (one to compare it to the Trees and another one to get an idea of the data)
- With: Accuracy = (TP + TN)/(TP + TN + FP + FN); Accuracy = ~2.6%
- With: Accuracy = (correct)/(total_values); Accuracy = ~73%

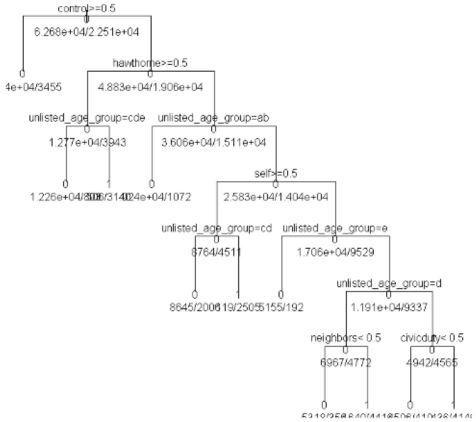
Part 3 – Plot the ROC curve for the model and compute the AUC of the model



-
- AUC = 0.568

Task 4 – Modeling with trees

Part 1- Using the rpart function in R and leaving all the parameters at their default values use the option method="class", build the model and plot the tree.



Part 1- Now, calculate the accuracy of the model. How does it compare to the logistic regression model?If better or worse, why?

- The accuracy of the tree model is of 87%.
- It is completely superior to the ~3% of the logistic regression data.
- This is because of the advantages of the decision tree model, as well as the disadvantages of the logistic regression model
  - The decision trees work specially well with data shaped like ours, the discrete categories allow for simple and efective branching. We must also note that not much data featuring is required to make useful decision trees.
  - Logistic regression does not work well with our data because it aims to return probabilities of outcome, and it only has categorical values. Another caveat is that we did not process our data further (eg. equalize the number of voters and non-voters)

Conclusion and next steps.

Conclusion.

Building this project has allowed us to understand the variables that affect whether a citizen will vote, and by grouping citizens in various categories we can now predict, with 87% accuracy, if a citizen will vote.

Next steps.

These results can be improved by further processing and analyzing the data, some possible paths can be:

- Improve our data metrics and tests, this way we can gather further insights on how the models are working and how we can improve them.
- Collect more data.
- Tune the models
- Improve our data featuring and include all our factors (here we did not include age)

Sources

- Some concepts and images were taken directly from websites, the sources are shown in the code (Jupyter notebook).