Income Prediction By Zip Code

Team Tutti Frutti
- Domenick Powers
- Kai Lau
- Min Jei Yeo
- Rommel Trejo

Background

The data provided by the MDP is organized by cities and encompasses numerous subjects including race, Income, Poverty, Employment and Labor data.

The Maryland Statistical Handbook, a publication by the MDP, for example does a great job by providing charts and graphs regarding the state of Maryland's different counties.

Problem Statement

The state of Maryland department of planning provides a variety of forecasts per region which depending on the publication a region may be a city, a county, or zip code.

Per region forecasts are useful when assessing the impact of policies that will affect an entire county or the entire state; nonetheless, in certain situations, it is more helpful to have per zip code forecasts: such as when opening a business.

The focus of this project shall be to provide financial and population forecasts per zip code from the Maryland statistical data provided by the Internal Revenue System (IRS).

Approach

- Our approach will be using the data collected from the previous years, we can predict the tax result with the implementation of linear regression.

Tools and Resources

- Coding Language: Python

- Tools for Data analysis:
  - Pandas
  - python
  - Matplotlib
  - Numpy

- Git and Github
- Linus Torvalds and Richard Stallman

Data

From IRS

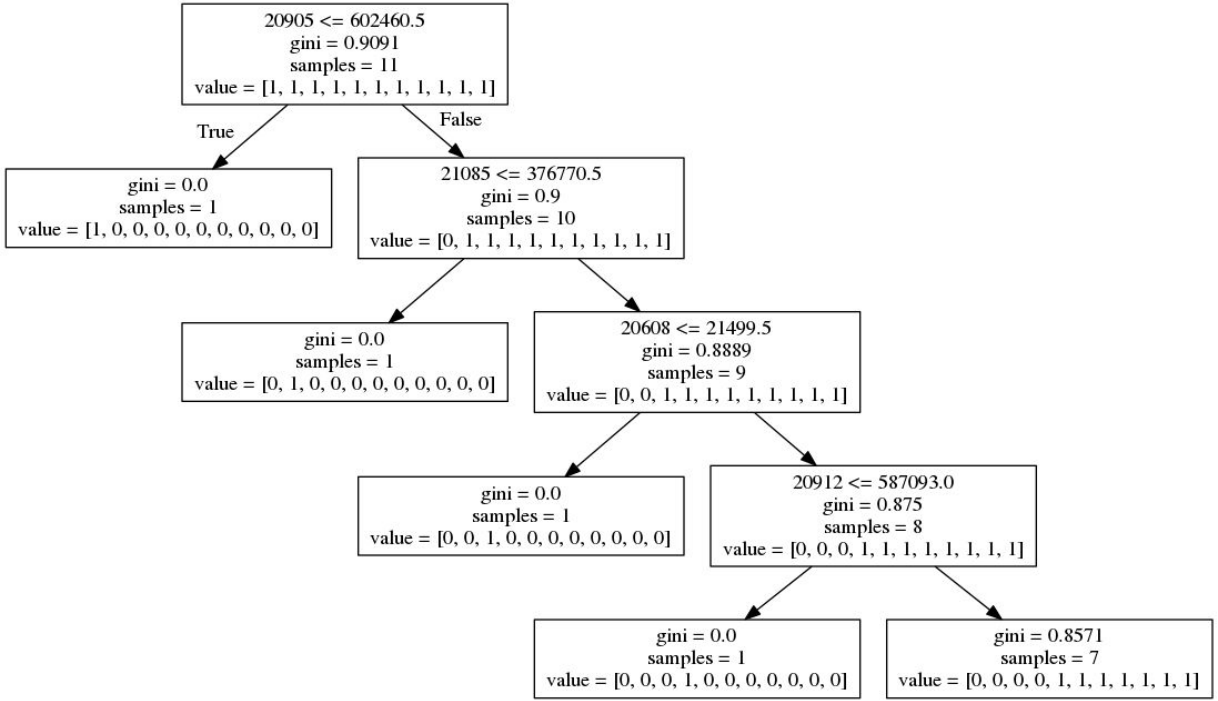
link to spreadsheet

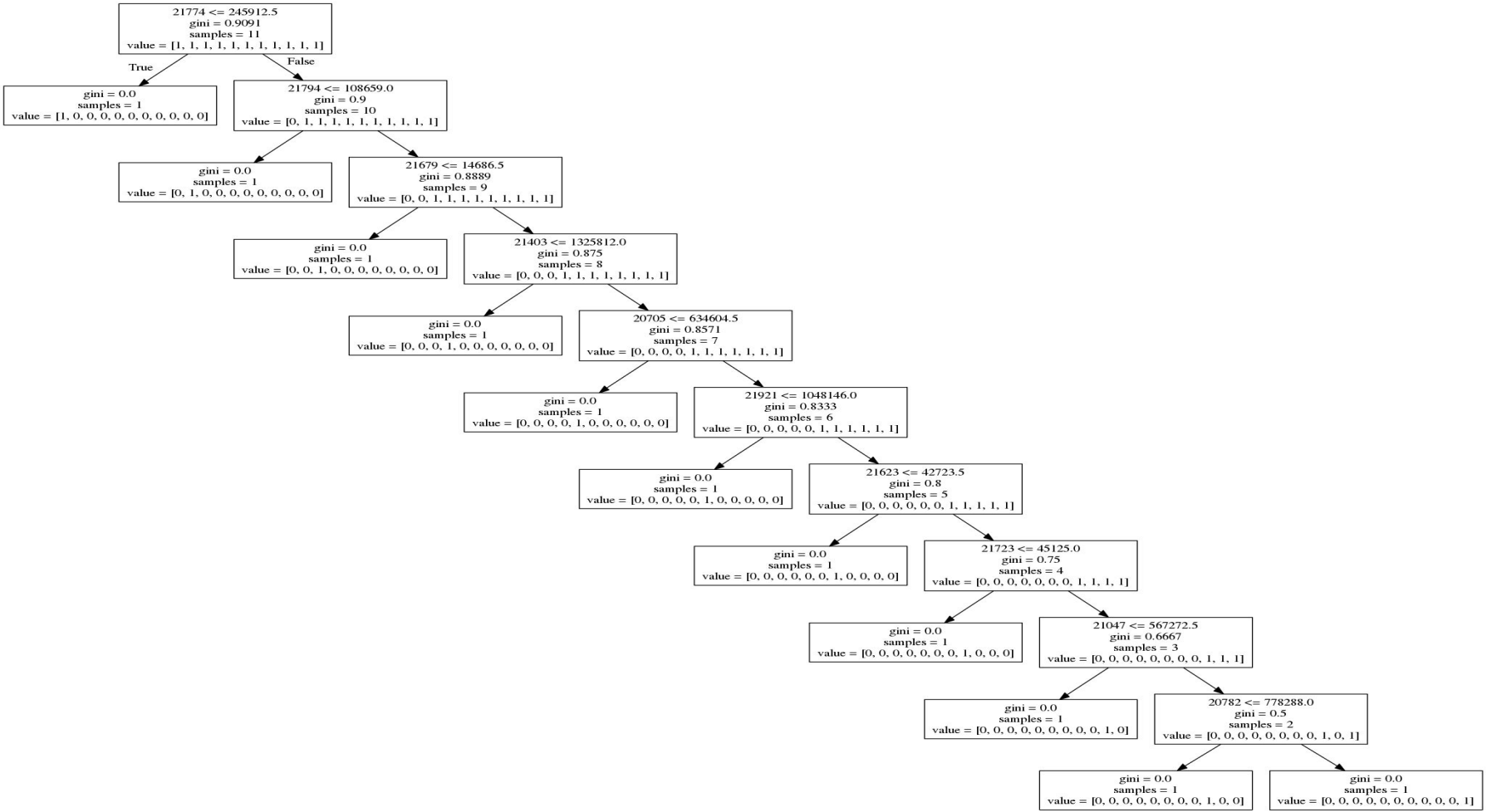
zip_code	Totals	20601	20602	20603	20606	20607	20608	20609	20611	20613	20615
1998	116614285	433388	348017	431102	6121	165562	15968	18323	21718	170758	7034
2001	136235786	528943	397740	563126	6872	207530	18336	22617	26359	208618	8506
2004	155654521	612853	462102	661844	7591	283660	21330	23921	34491	288261	12195
2005	166297995	645439	475835	702277	8517	301893	21669	25532	36202	316151	12201
2006	177136975	668446	509938	757255	7867	315711	25159	28695	31797	345948	13417
2007	190616115	703397	547553	819105	8675	346189	26349	31472	33469	371622	13726
2010	189470912	758051	639590	959395	9375	411954	26895	33827	35928	448973	13959
2011	196896143	768656	667192	988711	9799	434862	26729	32964	36699	475215	14344
2012	205576901	781967	686028	1013756	9224	448623	27064	34493	36236	487616	14210
2013	205257793	787861	695227	1044704	9282	460857	26294	33481	33992	501550	14464

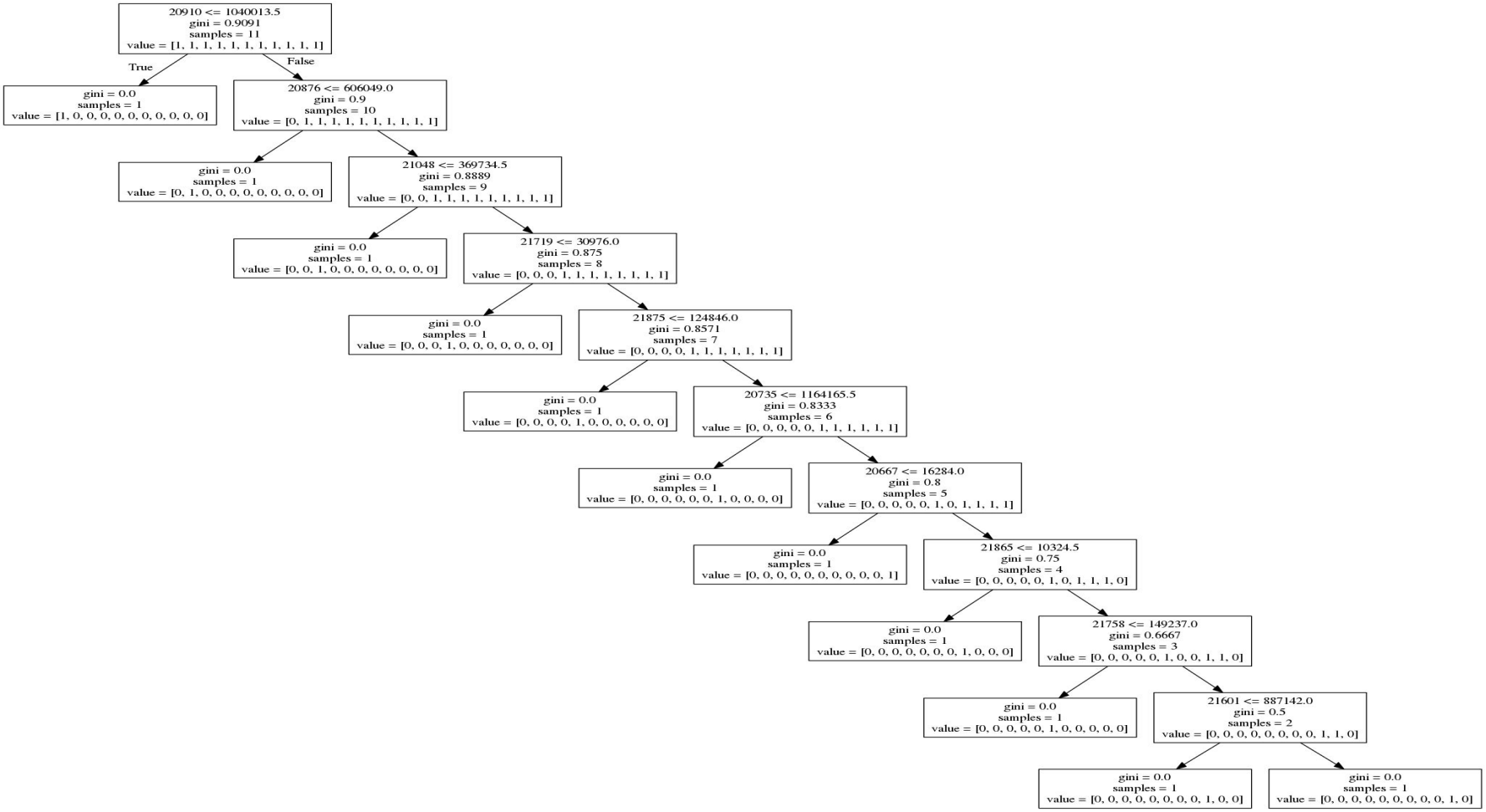
Algorithms

- Scikit-Learn : decision tree
- Support Vector Machine (SVM)
- Linear regression
- K means
- Neighbour search
  - 1
  - 10 *
  - 11

Demonstration




21774 <= 245912.5
gini = 0.9091
samples = 11
value = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
True
False
gini = 0.0
samples = 1
value = [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
21794 <= 108659.0
gini = 0.9
samples = 10
value = [0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
gini = 0.0
samples = 1
value = [0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]
21679 <= 14686.5
gini = 0.8889
samples = 9
value = [0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1]
gini = 0.0
samples = 1
value = [0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0]
21403 <= 1325812.0
gini = 0.875
samples = 8
value = [0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1]
gini = 0.0
samples = 1
value = [0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0]
20705 <= 634604.5
gini = 0.8571
samples = 7
value = [0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1]
gini = 0.0
samples = 1
value = [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0]
21921 <= 1048146.0
gini = 0.8333
samples = 6
value = [0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1]
gini = 0.0
samples = 1
value = [0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0]
21623 <= 42723.5
gini = 0.8
samples = 5
value = [0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1]
gini = 0.0
samples = 1
value = [0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0]
21723 <= 45125.0
gini = 0.75
samples = 4
value = [0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1]
gini = 0.0
samples = 1
value = [0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0]
21047 <= 567272.5
gini = 0.6667
samples = 3
value = [0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1]
gini = 0.0
samples = 1
value = [0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0]
20782 <= 778288.0
gini = 0.5
samples = 2
value = [0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1]
gini = 0.0
samples = 1
value = [0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0]
gini = 0.0
samples = 1
value = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1]


20910 <= 1040013.5
gini = 0.9091
samples = 11
value = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
True
False
gini = 0.0
samples = 1
value = [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
20876 <= 606049.0
gini = 0.9
samples = 10
value = [0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
gini = 0.0
samples = 1
value = [0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]
21048 <= 369734.5
gini = 0.8889
samples = 9
value = [0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1]
gini = 0.0
samples = 1
value = [0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0]
21719 <= 30976.0
gini = 0.875
samples = 8
value = [0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1]
gini = 0.0
samples = 1
value = [0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0]
21875 <= 124846.0
gini = 0.8571
samples = 7
value = [0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1]
gini = 0.0
samples = 1
value = [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0]
20735 <= 1164165.5
gini = 0.8333
samples = 6
value = [0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1]
gini = 0.0
samples = 1
value = [0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0]
20667 <= 16284.0
gini = 0.8
samples = 5
value = [0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1]
gini = 0.0
samples = 1
value = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1]
21865 <= 10324.5
gini = 0.75
samples = 4
value = [0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0]
gini = 0.0
samples = 1
value = [0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0]
21758 <= 149237.0
gini = 0.6667
samples = 3
value = [0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0]
gini = 0.0
samples = 1
value = [0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0]
21601 <= 887142.0
gini = 0.5
samples = 2
value = [0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0]
gini = 0.0
samples = 1
value = [0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0]
gini = 0.0
samples = 1
value = [0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0]

Issues

Dataset was continuous

- Algorithms overfitted
- Some years were missing
- Graphviz is …
    - Something else
        - everyone , probably

Improvements

Possible Changes:

- Decrease the sample size of our data.
- Use additional factors to create a better prediction

Conclusion

Impossible to predict where a person live based on income tax

Wealth of Maryland is increased by 60%

Reference

http://www.mdp.state.md.us/msdc/s3_projection.shtml

https://www.irs.gov/uac/soi-tax-stats-individual-income-tax-statistics-zip-code-data-soi

http://www.mdp.state.md.us/msdc/md_statistical_handbook.pdf