Calculating Cancer Incidents with Boosted Decision Trees

Daniel D. Meurer, MHA
Galvanize + ZNA

dandmeurer@gmail
github.com/DataDanD

Background

Our environment plays a very important role in our life and thus our health. According to the WHO, as much as 24% of all disease is caused by environmental exposures that could have been averted. [5]

40% of people will receive a diagnosis of cancer in their lifetime (½ of all men and ⅓ of all women). [4] Cancer rate is typically measured as incidents (new cases) per 100,000. Yearly California counties have an average of 540 incidents per 100,000.

Boosting models are go to base layer along with neural nets for most ensemble models in Kaggle competitions by Grand Masters. [6]

Questions

- *Why do cancer rates seem to differ widely in California countries?*
- *Can a supervised learning algorithm predict cancer rates at a county level?*
- *What model technique will give the lowest root mean squared error (RMSE)?*
- *What environmental factors will be important for predicting accuracy?*

Methods

- Find useful public datasets (any format)
- Filter data for appropriate features
- Group by year and county
- Pivot columns as needed
- Merge 14 dataframes to cancer data
- Feature engineering + more filtering
- Sk-Learn regression models & boosting
- Grid search parameters for lowest RMSE

Data

CA Data Sets	Group By	Rows	Columns	Features
SEER Cancer	YC	2,550,114	133	1
Population	YC	304,961	11	1
EPA Toxic	YC	73,979	113	2
EPA Air	YC	6,087	15	2
Infectious	YC	166,557	10	3
Radon	YC	1885	25	1
Fracking	C	3116	19	1
Superfunds	C	86	568	1
Health Status	C	58	28	3
Vulnerable Pop	C	58	28	3
Insurance	C	58	31	3
Education	C	58	9	1
Poverty	C	58	11	1
Employment	C	58	23	2
Stat FIPS	C	58	2	1

Y = Year | C = County

Datasets were collected from open source government data at data.gov. [1] Data is from surveys, reported by companies, or collected from government agencies. The columns and rows in the table above are recorded after filtering for California and for years 2001-2012.

Stack

python, ANACONDA, jupyter, aws, pandas, NumPy, CatBoost, scikit learn, matplotlib, +a|e|a|u

Results

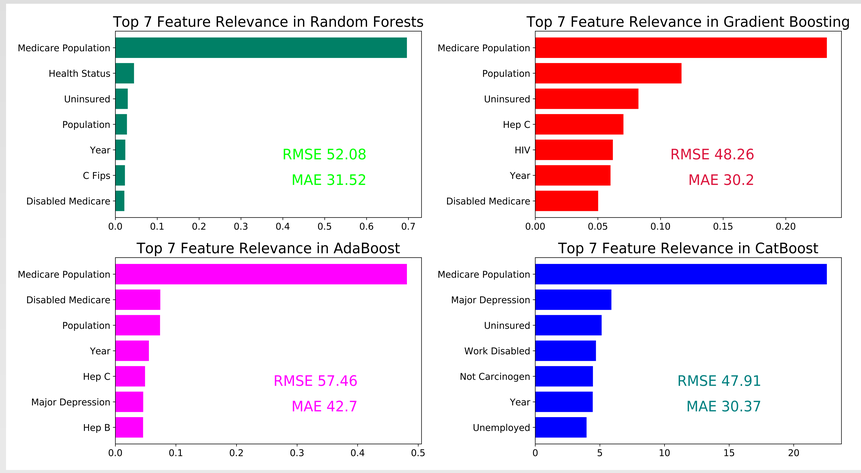

Figure 1. Feature relevance in boosted tree splitting

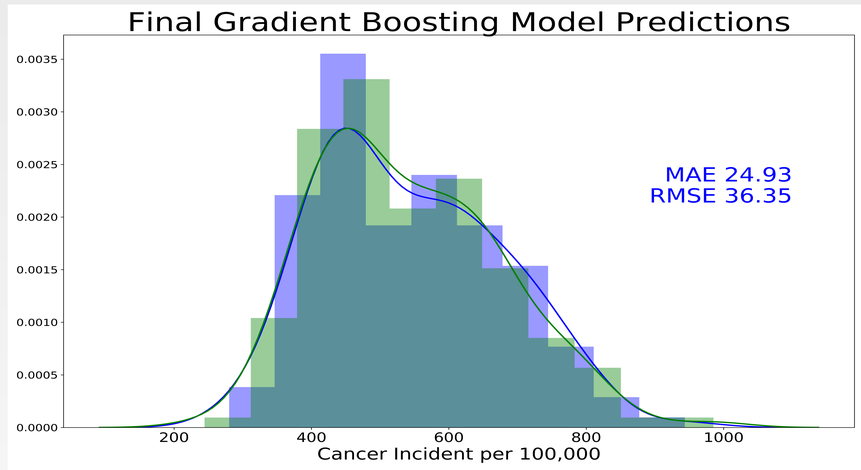

Figure 2. Final model predicting test data (BLUE) overlapped with actual values for that prediction (GREEN)

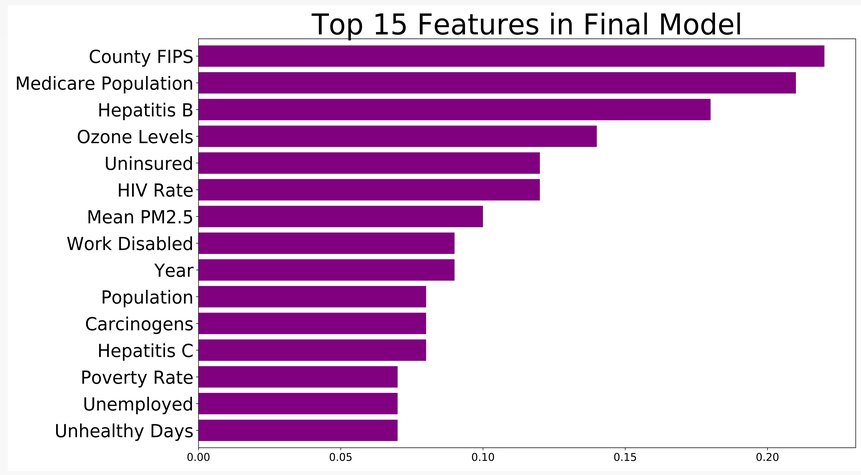

Figure 3. Model score change when feature is dropped

Discussion

This project was supposed to look at specific chemicals from the **Toxic Release Inventory** data from the EPA and specific cancers, but the data was sparse and not all toxins released into our environment have to be reported. Adding **socioeconomic** data made the model more accurate and reliable. The rows with the smaller counties could have been dropped, then it would have be possible to add more features: BMI, blood pressure, smoking. **Fracking wells** and **superfund sites** information would have performed better, if the cancer data was mapped to an area smaller than the county level. There may be some data leakage with **major depression**, as cancer can lead to depression. CatBoost / XGBoost / LightGB could have been good alternatives to gradient boosting, but they take longer to gridsearch for optimal parameters and implement.

Conclusion

Boosting works with this regression problem. The final model has variance of +/- 5 RMSE. **Medicare population** and **county FIPS** are the features that matter the most in predicting cancer incidence rates for these models. More data is needed to determine relationships between diseases and the environment.

References

- [1] https://www.data.gov/
- [2] https://seer.cancer.gov/data/
- [3] https://www.epa.gov/
- [4] https://www.cancer.org/cancer/cancer-causes.html
- [5] http://www.who.int/mediacentre/news/releases/2006/pr32/en/
- [6] http://blog.kaggle.com/2017/01/23/a-kaggle-master-explains-gradient-boosting/