

LOUMINI

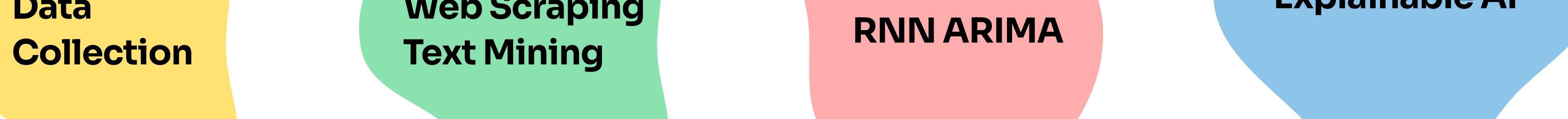
Investigating the effects
of air pollution on health
outcomes in the US

What is our Problem?

By the increasing public concern on the effects of air pollution on health how can US Government and political parties potentially use this as an opportunity to prioritize environmental policies and implement effective measures to address air pollution and its impact on health?

What is our Solution?

Using BA Tools to provide them an initial analysis where to build upon into more structured project



Collected data from 4 datasets for 39 US states, mainly from GOV.us & EPA.gov, ranging from 2000 to 2019:

- Health Indicator Dataset with 1.8 mil records
- Annual Air Pollution Dataset with 2K rows
- Demographical Dataset with 600K rows
- Tax information dataset of different states for 20 years

Managed to scrape 100000 articles from google news, by using 2 Keywords

▪ Managed to scrape 10k tweets
▪ Confirmed increase in awarness
▪ Got sentiment on Air Pollution and Cancer

Limitations

▪ Articles are not the full articles, 10k tweets are a small sample to get conclusions from

▪ More realistic results compared to ARIMA model

▪ managed to forecast 7 types of cancer based on US states for 6 years 2019-2025 by using more than 100 features including demographics, tax info and air pollutants by geographical location

▪ The MSE of RNN model calculated for Invasive Melanoma cancer is 0.13

• Getting a good overview of the most important features in our model using SHAP

• Information about how the value of each feature would affect the forecasted value

Limitations

• There are limitations to interpreting SHAP values as a universal explanation across all instances in the dataset due to differences in feature importance and contribution to the output of the model across different states and types of cancer.

Limitations

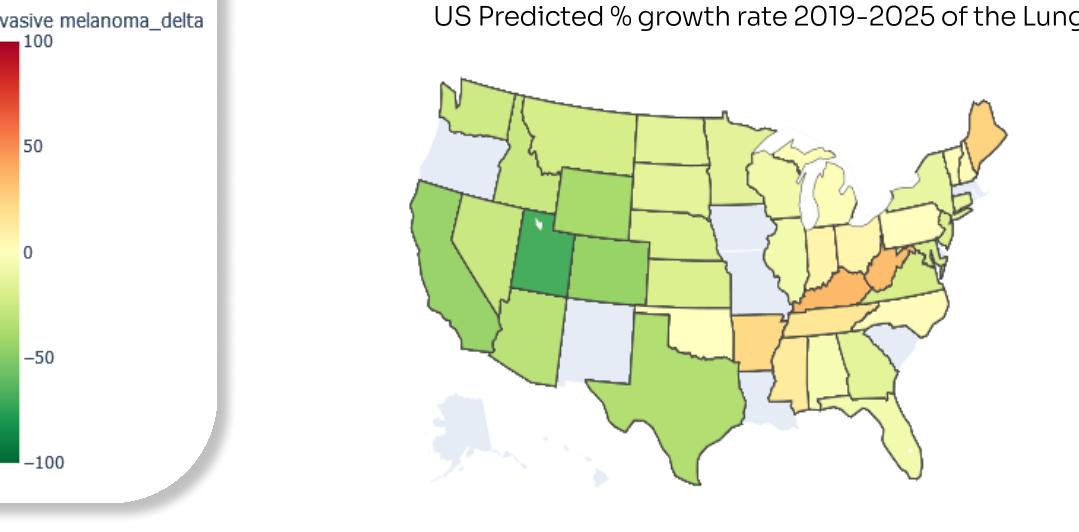
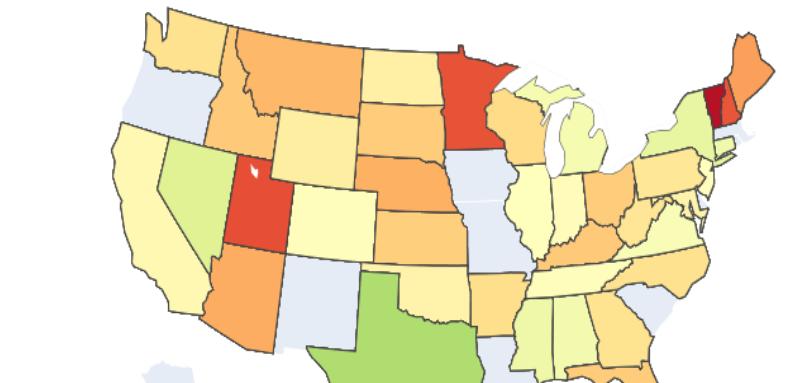
▪ Granularity of data is per each year & state. More concrete results could be derived with more detailed data.

Limitations

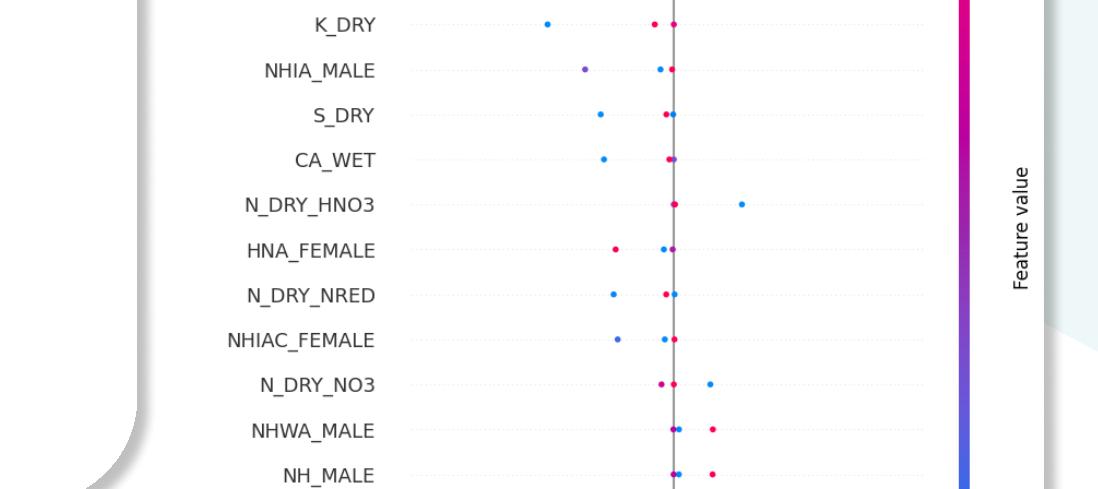
▪ Fitted RNN model on a specific state & cancer due to large combinations, with the assumption that it would apply to other states & cancers. Noteworthy that the model's accuracy may be lower for certain state and cancer combinations.

What are our Findings?

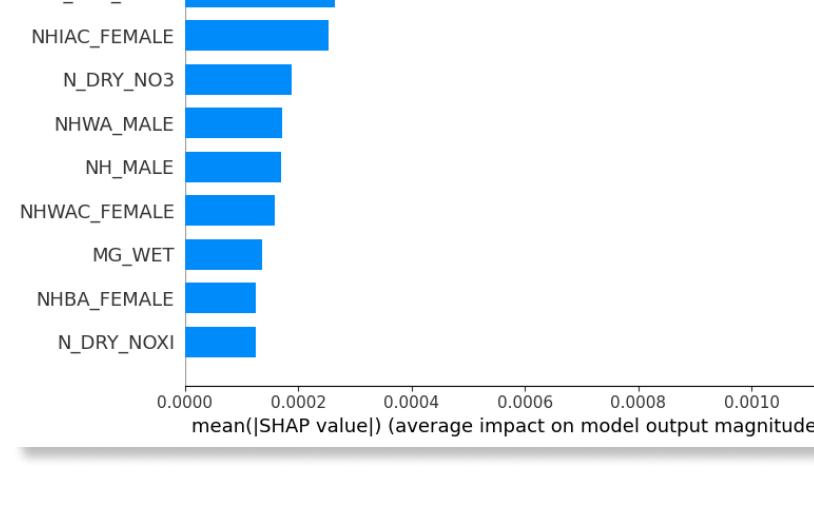
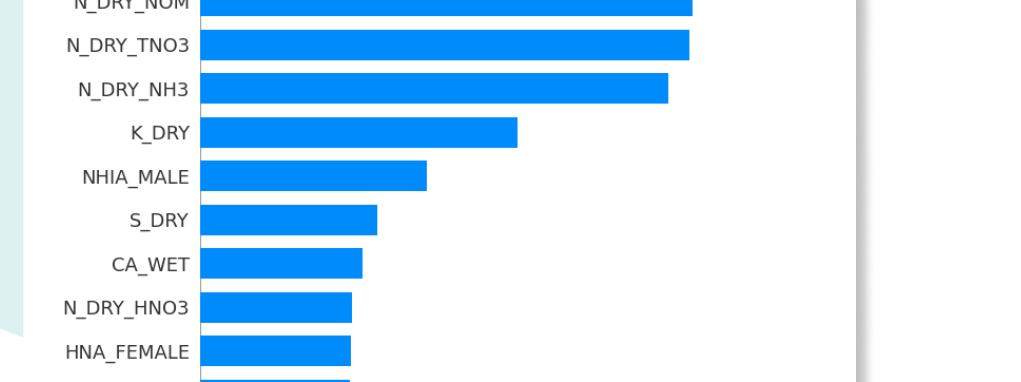
- Based on our RNN model, the most growth of Melanoma from 2020 to 2025 belongs to Vermont.
- As it can be seen in the SHAP plots, the value of "N_DRY_NH3NET" has the highest effect on the prediction of Melanoma cancer in Vermont, the amount of effect is also shown in the force plot.
- Increased coverage by the big News media outlets with concerned tones (sentiment analysis)
- The articles often highlight the connection between air pollution and its adverse effects on human health.



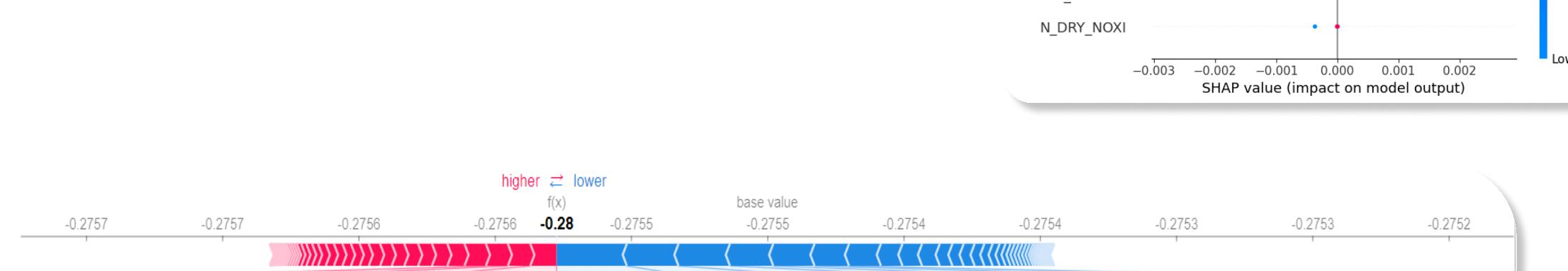
Combined Air Pollutants by State



Invasive melanoma % Delta from 2020 to 2025



US Predicted % growth rate 2019-2025 of the Lungs and bronchus' cancer



Now what's Next?

As our audience what to get from our project

Conclusions

While the provided information can support customer needs, it is limited. Consulting with health and environmental experts will be essential to verify assumptions before making any decision.

Next steps

This project did not aim to identify the specific causes of air pollutants. The next step could be to investigate and determine the true cause(s) behind the pollutants.