Barometric Pressure and Mood

Max Ramsdell

Github Repo: https://github.com/3lbsofSalt/DataScience8/tree/master
Slideshow: cs5830_final_project

Introduction

We all experience weather, and we all sometimes experience various mood shifts. What if they are connected? That's what we set out to figure out when we asked if there is a link between barometric pressure (and other weather data) and irritability. Our analysis will specifically focus on twitter sentiment analysis data in the country of Holland in the year 2020. This is important research, because research into the human condition is always important. Psychologists can use this understanding as a tool when helping people with erratic moods and everyday people can use this knowledge to help them understand themselves better.

Dataset

The data was actually significantly more difficult to procure than we initially thought. We chose Holland because that was the best twitter sentiment analysis data that we could find, and we had to search hard to find an appropriate weather data api. Once we were able to gather the data together it worked extremely well. We had time, date, and location data in both datasets, which allowed us to merge the twitter dataset with the weather dataset providing us with an extremely clean dataset ripe for analysis. It's also important to note here that the weather features we had access to were limited to barometric pressure, relative humidity, and surface solar radiation due to pricing constraints imposed by the api we used, which limited our analysis to some degree.

Analysis

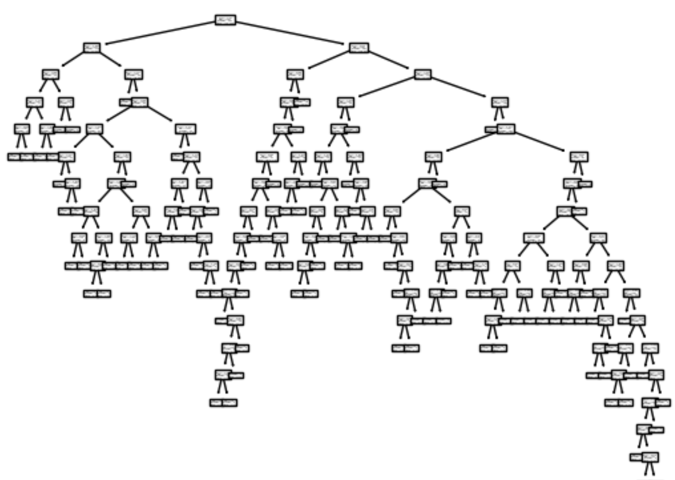
In this analysis we used a variety of techniques to try to understand the data. The main ones that were used were logistic regression, decision trees, and neural networks. We also attempted a linear regression and a k-nearest neighbors analysis to try to tease a relationship or correlation out of the data. Most of those models are classification models, and we used them to classify if a given day had an either positive or a negative mood. The linear regression worked because twitter sentiment data is represented as a scalar, allowing us to use linear regression as well. All of these models worked well because the data was conducive to them and each of them allowed us to view the data from a different perspective.

Results

Our analyses came back with inconclusive results. If our null hypothesis is that there is no relationship between a person's mood and the weather, then we were definitely not able to reject it.

Results from the linear regression were not good. None of the three weather values had a p-value less than 0.15, and ignoring the p-values, the correlation between the weather data and twitter sentiment were almost nothing. Logistic regression had a similar result, where every single model we generated had an all or nothing decision boundary causing it to classify everything as either likely to be positive or likely to be negative.

The results received from the decision tree and neural networks were a little better, but still not conclusive. For both, the negative cases did better than just guessing, with decision trees at an f1-score of 0.61, and the neural nets at an f1-score of 0.62. They both had reasonable precision and recall scores. For the positive cases, they did about or slightly better than just guessing. The problem with this is that we weren't able to learn anything from either of these models because the neural network is difficult to interpret and the decision tree was extremely complex as shown in the figure to the left.



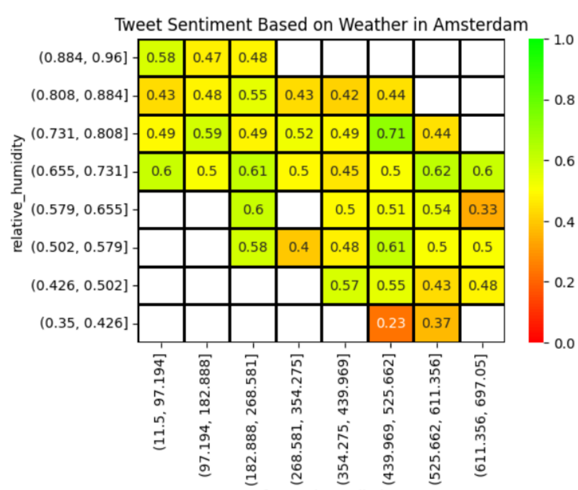
After this we attempted a k nearest neighbors classification as a last ditch effort. It performed surprisingly well, but we couldn't make sense of it either. It gave an f1-score on the negative case of 0.67, a drastic increase compared to the other models. We struggled to come up with an explanation for this. On one hand, we were looking for a very small effect size because someone might post a negative comment on twitter for any number of reasons, and if the negativity is affected by the weather, then the effect on that is likely to be tiny. These findings don't dispute that. On the other hand, our other models failed to show a distinct relationship. We have two hypotheses. Either our hypothesis is correct and the first few models performed poorly because they were not suited to the data, or the weather could help encode the time into the twitter data. If the time were encoded into it based on weather then the nearest neighbors are likely to be days that happened in the same week, and if there was a particular event that week, then perhaps many of the users were negative around the same time. At then end of it all, it's difficult data to interpret and these models mostly show that we cannot reject the null hypothesis yet, being mostly inconclusive analyses.

Technical

Preparing the dataset was one of the most difficult aspects of the project. Finding a good weather api that gave us accurate time data as well as good location data was difficult. When we did we had a limited number of api calls, forcing us to pick and choose what features we wanted, and we could only get two weather data points per day. Because of this we chose to get the weather at 10am and 4pm to provide a good spread.

We then needed to figure out which locations had the most tweets coming from them. We had to select only three cities due to api restrictions. This forced us to get rid of nearly two thirds of our dataset because the tweets came from "the Netherlands" (too general) or from smaller cities. Then we again halved our dataset by dropping all tweets that had a subjectivity score of 0. This was because if we are measuring mood, then we want the most subjective tweets as those will have mood encoded into them the strongest. This left us with about 15-20 thousand tweets to use. A significant amount, which we thought would be good enough. If we were to do this again with funding, it really would be nice to have tweet data and weather data from all across the world.

One more issue with our dataset was that it was from 2020. In that year lockdowns were in effect, with especially stringent rules in Holland as well. This could have meant that the weather wasn't having as much of an effect on people as it might have normally done.



As for the models, the first results we received came from the logistic regression analysis. We really fought with the model and in the end were unable to get a decision boundary out of it. It would only either mark all days as negative or all as positive depending on how we weighted the classes. We then attempted to perform a regression by binning the data, giving the logistic regression perhaps a simpler model to classify against. This performed the same as before. We then decided that logistic regression would not perform well. The graph above is informative as to why logistic regression did not perform well, and why we didn't try a support vector machine even though we initially planned on it. The graph is fairly similar to other graphs we generated with different combinations of weather data.

When we attempted to model the data using the neural network and decision tree models, the models seemed to perform very well because they are better at modeling complex relationships than logistic regressions. However, the models became so complex we struggled to get any knowledge out of them.

The k-nearest neighbors model is a special case. We actually didn't think that it was a good model for our data, because of the previously mentioned lack of patterns but decided to try it anyway. It performed so much better than the other ones (specifically for the negative case) that we couldn't figure out how to explain it. Looking at the chart above again, there do seem to be clusters of data, but those clusters don't seem to have a pattern of any kind. Perhaps a weather scientist, or better yet, more related weather data would help us to explain what seems like an anomaly, but as it stands, we cannot come up with a satisfying explanation for it.