Understanding Influenza Outbreaks Using Historical Data and Socioeconomic Factors

Andrei Kapustin - ak7671,
Shiva Sanketh Ramagiri Mathad - srm714,
Avinash Prabhuling Veershetty - apv280,
Roman Zamishka - rpz205

Business Understanding

Despite being a common annual occurrence, the influenza virus, also known simply as the flu, is one of the leading causes of death in the United States and around the world. By depressing the immune system of the host, especially in vulnerable children and elderly, influenza paves the way for more dangerous lung infections to cause pneumonia. In 2017 influenza was associated with 55,672 deaths in the U.S., making it the 8th leading cause of death that year.[1] Although flu vaccines are an important tool available to reduce the public health impact of influenza, because of the slow process with which these vaccines are manufactured they have to be made months ahead of the flu season. The long lead times force manufacturers to develop vaccines based on an estimate of the influenza virus strain that will be dominant in the upcoming season, which can result in the vaccine being less effective if the guess is wrong. Because of this high uncertainty, the Center for Disease Control (CDC) has made it a national priority to track the spread of influenza by maintaining a weekly national database of key metrics related to influenza. The two most important of these metrics are the influenza-like-illnesses (ILI), which tracks the number of patients who presented themselves to the healthcare system with symptoms of influenza, and a count of the influenza cases that were confirmed by laboratory tests. The goal of the CDC is to use this data to better predict the spread of the flu in order to make better plans about resource allocation and public engagement in the vaccination campaign with the aim of reducing the

number of infections and deaths. Although these metrics are reported on a weekly basis, there is a one week lag period between hospitals collecting the data and the results being reported because the CDC requires a week to aggregate and analyze the data that they've received. It is a critical mission for researchers to fill this knowledge gap with an effective forecasting model because the incubation period for the flu is an average 2 days.[2] Currently the CDC issues notices about elevated flu seasons, but if the CDC were able to forecast a surge in influenza cases in advance, it would be able to issue an early warning to the healthcare system to prepare their emergency rooms by stocking supplies and maintaining sufficient staffing, as well as increase their public vaccination awareness efforts. Previously attempts have been made to improve the forecasting model with weather data, as temperature and precipitation are important to the flu incubation process.[3] In addition, data mining has been used to improve the model with the inclusion of search and social media data.[3,4,5,6] However, many of these attempts are at the location-insensitive national level or only focus on one county. In addition, recent health policy research has shown that economic and social factors are important determinants of health[7,8]. We develop a model for all NY state counties and attempt to improve the model with a unique recently-released county-level economic dataset.

Data Understanding

The CDC publishes its flu surveillance data and maintains an interactive visualization tool on its FluView website[9]. The data available publicly on FluView, however, is aggregated to the state level which makes it less precise than data available at the county level. The primary database used is a county level database specific to the State of New York for the period 2009 and 2019 reporting the number of laboratory-confirmed influenza cases.[10] The target variable is

chosen to be laboratory-confirmed influenza cases over the ILI metric because only 10-20% of people who visit the doctor for the flu actually have the infection[11] and the ILI contains noise from other seasonal diseases such as the common cold.

The Laboratory-Confirmed Influenza Case dataset contains 20,080 county/week pairs and provides the data for the county, the weekend date, the flu season week as defined by the CDC, and the flu counts for that period broken down by the type of flu (A,B, or unspecified).

Weather is an important factor in the spread of the flu because the virus spreads most effectively in cold and dry seasons.[12] The National Oceanic and Atmospheric Administration provides local weather data through its regional climate centers (RCC). Data for New York state counties is sourced from the Northeast RCC's Climod system.[13] The system allows to search for a city or county and provides a map of nearby weather stations. The station chosen for each county is the station closest to the county centroid and without any significant gaps in the data. The weather dataset provides daily max temperature, minimum temperature, average temperature, precipitation, snowfall, and snow depth.

Google Trends provides relative search popularity for various search terms on a weekly and regional basis.[14] Notably the Google Trend reports on a weekly basis for data pulls that contain a maximum of 5 years, and reports on a monthly basis for longer periods. Data is pulled in three pulls for periods 2009-2011, 2012-2015, and 2016-2019 for the search terms "flu", "flu symptoms", "fever", "cough", and "sore throat".

The CDC collects and provides data on the FluVaxView site on the vaccination rate estimates from the Behavioral Risk Factor Surveillance System (BRFSS) and National

Immunization Survey (NIS).[15] The FluVaxView data contains monthly state-level estimates for vaccination rates for adults and children for the period 2013-2018, but because only the child data is differentiated between New York City and the rest of the state, only this data is used in the model.

Income and employment are important social determinants of health, making economic activity a novel addition to influenza modeling. Although economic data is normally reported at the state level, the Bureau for Economic Analysis has released county level GDP data from which GDP and GDP change data are added to the model.[16] In addition, the United States Department of Agriculture provides annual county-level demographic data in the Atlas of Rural and Small-Town America.[17] This data source is used for population level and unemployment level estimates for the model.

Data Bias: The flu vaccination data is based on telephone surveys and could potentially not reflect the general population. The GDP data is based on a trial dataset released by the BEA and may contain errors not recognized at the time of its release. In addition the GDP and Atlas of Rural and Small America datasets have annual data instead of weekly. Applying this data in a linear fashion across the year may introduce a bias to the model.

Data Preparation

The target variable being forecast in this model is chosen to be the total number of laboratory-confirmed influenza cases in a county in two weeks from the time of the most recent report. This means that, for example, as the CDC is publishing the final analysis for week 30 in week 31, the model is estimating the number of confirmed cases in week 32. Although this limits

the amount of public health impact that can be gained from this early warning, the CDC also has a responsibility to maintain the public's confidence in the institution. Recent trends of resistance to vaccination in the anti-vax movement could be emboldened by inaccurate CDC forecasts, and longer forecast horizons would decrease the accuracy of the model. The 2-week forecast (1-week warning) target decreases this risk and is still enough time to have some impact against the 2-day influenza incubation period, as well as allow hospitals and other providers to prepare for a surge.

Before the datasets could be merged many had to be reformatted to fit the county/week index used in the primary Laboratory-Confirmed Influenza Cases dataset.

The Laboratory-Confirmed Influenza dataset reports each county/week pair in three lines corresponding to the type of influenza that was confirmed (A, B, unspecified). These lines are merged to obtain a single aggregate count, and the influenza type counts are converted into feature columns.

The weather data comes in the format of daily reports. This data is merged to match the weeks in the Influenza dataset and the average of the merged days is used as the value of the new data.

The Google Trend data must be downloaded in periods spanning no more than 5 years. This causes the popularity rank data to normalize relative to the highest search popularity in that period, with that data point representing 100. These years have to be normalized back to being relative to the highest popularity week over the full 2009-2019 period being analyzed. In addition, Google Trends data is reported at the metropolitan area level instead of county. Because each

metropolitan area includes multiple counties, an analysis of these areas is completed to sort each county into a metro area and then each entry is matched to the appropriate county.

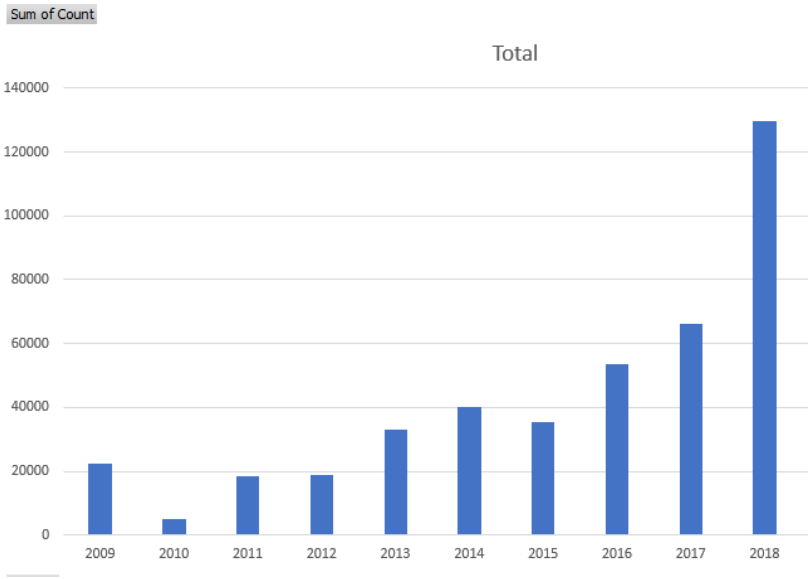
The data for year 2019 is dropped from the analysis because this year is not yet complete and many of the data sets did not report on this year. Missing weather values are replaced with the average value for that county for that week across all years ie. a missing value in week 48 is replaced with the average of the values in the other weeks 48 for that county across all years in the dataset. Missing values for GDP and population are filled in with the most recent available value.

Data Analysis

Visual analysis of the reported Influenza cases and the 'SNOW' data (appendix) confirms that these features are positively correlated. Peak values for both are between the last 3 weeks of the year and the first 10 weeks of the next year, and both drop to less than 20% between weeks 17 and 42. We also conclude that the confirmed Influenza cases and 'TAVG' (Average temperature) are negatively correlated (appendix). From these observations we can confirm that the spread of Influenza is supported by cold weather conditions as expected.

Because the population in a large county may overstate the level of interaction between people, a county population density feature is engineered. In addition, it is important to note that the future weather will not be known when forecasting, therefore future weather features are generated based on the average weather for that month in the dataset, similarly to how the missing weather data was filled. Both the weather and the forecasted weather are kept in the forecasting model because while the future weather features provide an estimate of the

weather, the current weather features provide information on how the current season deviates from the historical average. Further analysis of the weather data shows that max temperature and min temperature closely track the average temperature throughout the year (Appendix) and are dropped. We choose to not encode the county name into dummy variables because this would result in 63 additional features without providing much more information over what is already contained in the demographics and economic data. Analysis of the influenza confirmed-cases counts reveals that the data is unbalanced and that 2018 had far more cases than the preceding years. This is consistent with our knowledge that 2018 was the worst flu season in 40 years.[18] We choose to use the 2045 data points in the Year 2018 data for the test set because being able to predict an extreme flu season would be an important measure of the model's performance. The remaining data is split randomly into a 16,484 datapoint training set and 2046 datapoint validation set.



Modeling and Evaluation

Linear Regression: Because the model is trying to predict continuous variable counts in a specific week we focus on regression models. Linear regression with few features is chosen as the base model for its simplicity and common use with continuous data. We choose the R^2 as the evaluation metric of the models for its high comprehensibility and ability to be used effectively across different models. The related Mean-Square-Error (MSE) is not chosen because the range

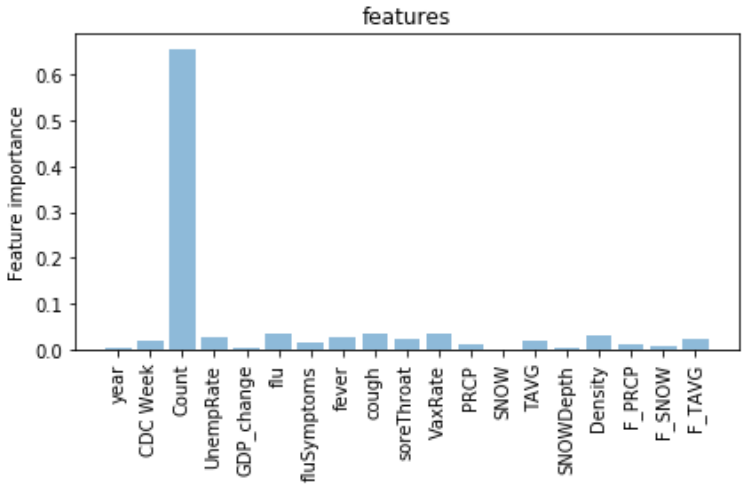
of the target feature [0,2401] would result in very high MSE values that would be more difficult to interpret. In addition the R^2 is chosen over the AIC because the total feature count is relatively low and we are focused on maximizing predictive power. R^2 is also directly related to the CDC's need to explain variation with few features. The base linear regression model including the ['year', 'Count','CDC Week', 'flu' ,'fluSymptoms', 'fever', 'cough', 'soreThroat','VaxRate', 'PRCP', 'SNOW', 'TAVG', 'SNOWDepth', 'Density'] features produces a baseline R^2 of .656. The lasso regression is inappropriate for this dataset because the relatively low feature count and the importance of a few key features such as 'Count', representing the most recent count of lab-confirmed flu cases, would result in features being eliminated. In order to improve the model a ridge regression is attempted for its ability to deal with multicollinearity, but produces a worse R^2 of .655. Adding the remaining 'GDP','GDP_Change','UnempRate' and 'F_PRCP','F_Snow', and 'F_TAVG' features only improves the linear regression model R^2 to .657.

Stochastic Gradient Descent: Stochastic Gradient Descent (SGD) is an iterative optimization algorithm that tests different regression coefficient values while trying to minimize their cost function. The SGD regression of the full model provides no improvement over the base linear regression, with a R^2 of .651. This is possibly because the SGD performance gains relative to closed form equations are realized with much larger datasets.

Support Vector Regression: Although the Support Vector Machine (SVM) is commonly used for classification tasks, it is also possible to use it as a regression (SVR) strategy. Unlike simple regression which tries to decrease the error rate, SVR is trying to determine an error rate within a decision boundary distance from the hyperplane that maximizes the number of data points

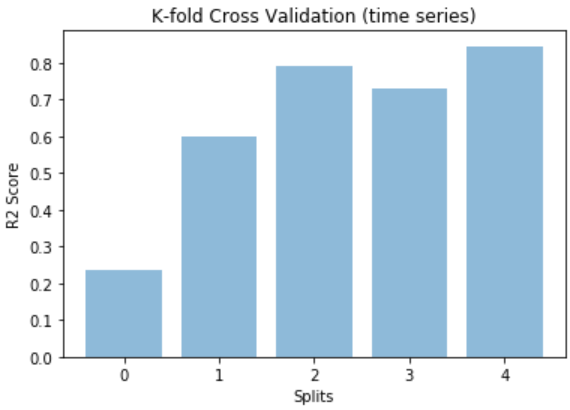
within that boundary. The SVR performs very poorly with R^2 of -.44, indicating that the model performs worse than a linear regression set to the mean.

Decision Tree Regression: The decision tree seeks to develop a regression by building a tree structure representing association between predictive variables. This is a powerful strategy for this dataset because we know that there are relationships between the features that result in higher flu counts, such as low precipitation and low temperature which promotes virus incubation, or high count and low vaccination rate which accelerates the spread of the virus. The default parameter decision tree results in R^2 of 1.0 on the training set but only .63 on the validation set. The much lower R^2 score on the validation set indicates over-tuning and that the model suffers from high variance. Tuning parameters for leaf size and minimum split size using MSE as the decision criterion for the quality of a split results in a min sample leaf value=4096 and min split value=2 with a train R^2 score of .80 and validation R^2 score of .70 . In addition to these results, feature importance analysis reveals that the tree is heavily reliant on the most recent flu vaccine count. While this is expected, the other features still contribute approximately a third of the model's explanatory power. Other important features are the population density, the vaccination rate, the 'cough' and the 'flu' Google Trend terms, and the seasonal average temperature.



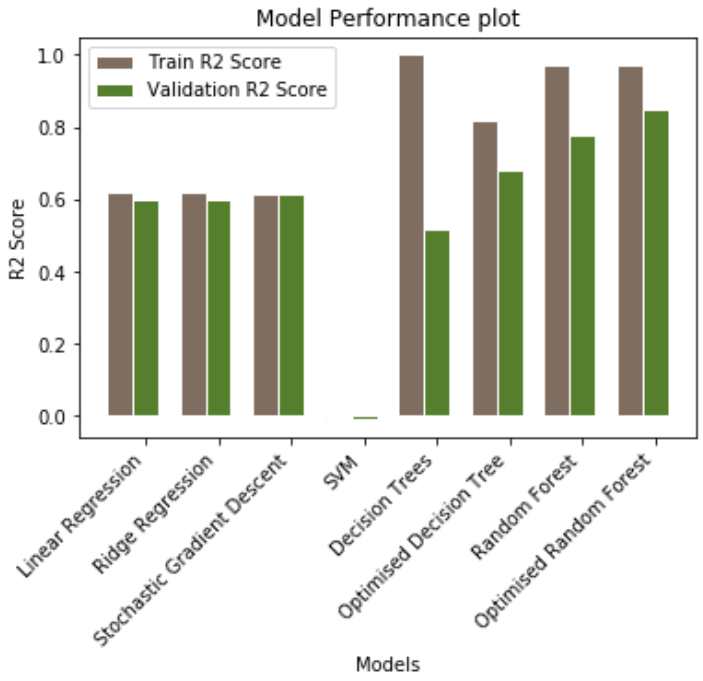
Random Forest: The decision tree provides the first successful improvement in forecasting over the base linear regression model. The decision tree results are further improved by developing a random forest ensemble (RF) model. The RF seeks to improve the model by bootstrapping many different samples from the dataset with each sample unique from the original data, then fitting the samples and aggregating the results. Single decision trees are vulnerable to overfitting the data and result in high model variance, which limits their generalizability, but the average of the predictions of the estimator trees in RF corrects for this and should result in a reduction of the variance. The base random forest model with 500 estimator trees provides a meaningful improvement over the decision tree model with a training R^2 score of .97 and validation R^2 score of .81. Further parameter tuning for the number of estimator trees determines that the optimal number is 900 trees, and results in a training R^2 score of .97 and validation R^2 score of .82 .

K-Fold Cross-Validation: K-Fold Cross Validation (CV) is a resampling strategy used to increase the robustness of models trained on limited datasets with the goal of increasing the robustness of the model. CV divides the data into k 'folds' and draws samples of k-1 folds with the k fold serves as a testing holdout, until all folds have been tested. Although CV partially duplicates the sampling process in random forest bagging, it is attempted here because of the small data size. Because the model is forecasting a time series value it is important to retain the information of the data's order and a CV TimeSeriesSplit is used instead of the traditional k-fold



CV. After parameter tuning it is determined that the best number of splits=5, resulting in a training R^2 score of .87 and validation R^2 score of .97

Model Selection and Testing: The best model is identified as the Random Forest with 900 estimators and time series cross validation. However, testing this final model against the holdout 2018 test data results in a R^2 score of only .638 compared to the base linear regression model's R^2 score of .637 . It's necessary to highlight again that 2018 was the worst flu season in 40 years and had double the number of confirmed flu cases of previous years. In the feature importance analysis we noted that the model was heavily reliant on the most recent recorded case count. This feature is the only one in this dataset that captures the virulence of the flu and we believe that in an extreme season the importance of this feature will dominate the associations between features identified in the random forest. To test this theory we drop 2018 from the dataset and retrain the models on years 2009-2016 with the year 2017 data serving as the holdout test set. As predicted, this results in improvement of the optimized RF model's performance with a R^2 score of .77 on the test data.



Deployment

The model is able to forecast the number of laboratory confirmed cases to a reasonable degree during normal flu seasons, but performs poorly at predicting 'black swan' seasons with very virulent strains of influenza. This is a known issue as Google's own forecasting model Google Flu Trends was shuttered because of its inability to detect major surges.[19] Because of the model's low R^2 performance during extreme flu seasons, situations in which it would be expected by the public to be the most useful, the model should not be used for making public forecasts. The model can still be used internally during normal flu seasons for making decisions about vaccination public engagement or making recommendations to hospitals. In addition, even during extreme seasons, the model's sudden poor performance can be used as a signal to the CDC that a particularly virulent flu season is approaching. Much of the data is gathered from surveys and datasets collected by the various federal government agencies and a data pipeline should be constructed allowing this data to come to the CDC directly instead of manually gathered. It must be noted that the model must be retrained frequently, as its possible for the forecast to be correct but appear too high because the interventions it triggered reduced the actual flu infection count. Frequent AB testing and other further research would be needed to tune for this effect. Further research would also be necessary to identify features better capable of modeling the virulence of the season's flu. In addition, social determinant factors like GDP and unemployment were not important features. This may be because they were reported on an annual basis instead of weekly, and obtaining more granular data may result in better performance. Finally, flu vaccination was a relatively valuable feature, but this data is not readily available on a county basis. The CDC should consider enhancing its surveying and

coordinating data with common vaccination locations such as schools and hospitals to improve the strength of this feature.

Ethical Considerations

The model uses entirely anonymized aggregated data, so there is no confidentiality. However, demographic values such as racial composition were not included in the model because they are not available in sufficiently granular form. This means that the forecast model could unintentionally result in disparate impact on specific groups. In addition, if the forecast model is used to make recommendations to hospitals then it must be extremely sensitive to extreme seasons as any overallocation of resources of flu prevention would mean other clinical areas are being put on the sideline, while under-allocation due to reaction to a too-low forecast would result in unnecessary deaths associated with the flu.

Bibliography

1. Nichols, Hannah. "The Top 10 Leading Causes of Death in the United States." *Medical News Today*, MediLexicon International, 4 July 2019, www.medicalnewstoday.com/articles/282929.php#heart-disease.
2. *Flu Facts: Incubation Period and When It's Contagious*. Healthline, 28 Oct. 2018, www.healthline.com/health/flu-incubation-period.
3. Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS (2015) Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. PLoS Comput Biol 11(10): e1004513. doi:10.1371/journal.pcbi.1004513
4. Dugas AF, Jalalpour M, Gel Y, Levin S, Torcaso F, et al. (2013) Influenza Forecasting with Google Flu Trends. PLoS ONE 8(2): e56176. doi:10.1371/ journal.pone.0056176
5. Cécile Viboud, Pierre-Yves Boëlle, Fabrice Carrat, Alain-Jacques Valleron, Antoine Flahault (2003) Prediction of the Spread of Influenza Epidemics by the Method of Analogues. American Journal of Epidemiology, Volume 158, Issue 10, Pages 996–1006,
6. Doornik, J.A. (2010). Improving the Timeliness of Data on Influenza-like Illnesses using Google Search Data.
7. Len M. Nichols and Lauren A. Taylor (2018) Social Determinants As Public Goods: A New Approach To Financing Key Investments In Healthy Communities. Health Affairs 37:8, 1223-1230

8. Epstein, D., Jiménez-Rubio, D., Smith, P.C. and Suhrcke, M. (2009), Social determinants of health: an economic perspective. Health Econ., 18: 495-502. doi:10.1002/hec.1490
9. "Weekly U.S. Influenza Surveillance Report." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 5 Nov. 2019, www.cdc.gov/flu/weekly/index.htm.
10. "Influenza Laboratory-Confirmed Cases By County: Beginning 2009-10 Season." *Influenza Laboratory-Confirmed Cases By County: Beginning 2009-10 Season | HealthData.gov*, healthdata.gov/dataset/influenza-laboratory-confirmed-cases-county-beginning-2009-10-season.
11. Salzberg, Steven. "Why Google Flu Is A Failure." *Forbes*, Forbes Magazine, 23 Mar. 2014, www.forbes.com/sites/stevensalzberg/2014/03/23/why-google-flu-is-a-failure/#2c4c95c75535.
12. Sommerville, Annmarie, et al. "The Reason for the Season: Why Flu Strikes in Winter." *Science in the News*, 4 Dec. 2016, sitn.hms.harvard.edu/flash/2014/the-reason-for-the-season-why-flu-strikes-in-winter/
13. "CLIMOD 2." *CLIMOD 2*, 12 Nov. 2019, climod2.nrcc.cornell.edu/.
14. "Google Trends" 12 Nov. 2019 https://trends.google.com/trends/explore?geo=US-NY&q=flu
15. "2018-19 Flu Season." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 8 Oct. 2019, www.cdc.gov/flu/fluvaxview/1819season.htm.
16. "GDP by County." *U.S. Bureau of Economic Analysis (BEA)*, www.bea.gov/data/gdp/gdp-county.
17. "Atlas-of-Rural-and-Small-Town-America." *USDA ERS - Download the Data*, www.ers.usda.gov/data-products/atlas-of-rural-and-small-town-america/download-the-data/
18. Lena Sun, Lindsey Bever. "This Flu Season Has Now Reached Pandemic Levels (but It's Not Technically a Pandemic)." *The Washington Post*, WP Company, 30 Mar. 2019, www.washingtonpost.com/news/to-your-health/wp/2018/02/09/this-flu-season-has-now-reached-pandemic-levels-but-its-not-technically-a-pandemic/.
19. Kennedy, David Lazer and Ryan. "What We Can Learn From the Epic Failure of Google Flu Trends." *Wired*, Conde Nast, 6 June 2015, www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/.

Appendix

All code and data for the project can be found in the project GitHub

Contributions:

Andrei Kapustin - literature review, data cleaning, feature engineering, deep learning model (dropped), report proofreading

Shiva Sanketh Ramagiri Mathad - data cleaning, feature engineering, base model, SGD, SVR, DT, FR, report writeup

Avinash Prabhuling Veershetty - data cleaning, feature engineering, SGD, SVR, DT, FR, report proofreading

Roman Zamishka - literature review, business and data understanding, data cleaning, feature engineering, base model, report writeup

Weather Feature Variation over the Year