Predicting Influenza Infection Chances and Identifying Risk Factors

Kaung Khin kkhin/kkhin@andrew.cmu.edu
Heinz College
Carnegie Mellon University
Pittsburgh, PA, United States

Shawn Leahy sleahy/sleahy@andrew.cmu.edu
Heinz College
Carnegie Mellon University
Pittsburgh, PA, United States

Abstract

Influenza is a highly contagious disease prone to cause epidemics around the world. Predicting individuals with a high chance of infection and identifying the spread of the disease could allow health care systems to better optimize how resources are used to combat the spread. Predicting outbreaks is a multidimensional problem involving many different fields. In this paper we investigate how historical medical and demographic survey information combined with Internet search engine use can be used to predict the chance of infection for individuals in different regions of the United States. We compare the effectiveness of the Random Forest Algorithm with Gradient Boosted Trees and evaluate using ROC curves and the AUC metric.

1. Introduction

Influenza, or simply the flu, is a seasonal contagious respiratory disease that affects millions of people in the United States yearly with outcomes ranging from just mild symptoms to even death. (CDC, 2017) It is caused by the influenza virus with the most infamous version in recent time being the 1918 flu pandemic nicknamed the Spanish Flu. (Rolfes et al., 2017) Although the disease is commonly self-limiting, it can progress to influenza pneumonia, which has a significant mortality. Up to 50,000 are killed each year by influenza-like illnesses. Because influenza causes significant morbidity, can be fatal, and often presents with new strains, the prediction of vulnerable populations can be of value in allowing for timely preventive public health planning and interventions to be used to mitigate the effect of these outbreaks.

The economic cost of the flu on the United States is estimated to be in the billions of dollars. Questions such as which states or counties bear high costs and where to distribute vaccines to achieve the maximum returns are crucial to improving future effectiveness of the health care system.

In 2008 Google explored flu forecasting in real time based on search histories. Unfortunately, at the peak of the 2013 flu season, the predictions were off by more than 140 percent. (Lazer et al., 2014) The use of big data to track flu trends is not a novel idea, in fact many studies have attempted to learn from the mistakes of Google Flu Trends. The challenges of using big data to attempt this type of prediction is well-documented. (Lazer et al., 2014) However, there has been some success in the use of social media data and search

data combined with health surveillance data to track and predict the number of cases for the Zika virus outbreak in Latin America. (Mcgough et al., 2017)

Here we demonstrate that the tracking of influenza search engine tracking combined with demographic and medical data and a suitable machine learning algorithm can accurately predict influenza infection probabilities (Ginsberg et al., 2009) and assist in identifying at risk populations at the regional scale in the United States.

2. Methods

Data

We utilize survey data collected from the Medical Expenditure Panel Survey (MEPS) from the Agency for Healthcare Research and Quality (AHRQ) a section of the U.S. Department of Health and Human Services. MEPS is a set of large-scale surveys of families and individuals, their medical providers (doctors, hospitals, pharmacies, etc.), and employers across the United States. The Household Component collects data from a sample of families and individuals in selected communities across the United States, drawn from a nationally representative subsample. The Medical Conditions file provides information on household-reported medical conditions. The Prescribed Medicines file contains household-reported prescribed medicine that was purchased during calendar year. (AHRQ)

Table 1: Unique Individuals in Each Survey Year

Year	Sample Size
2015	35,427
2014	34,875
2013	36,940
2012	38,974

Google trends is a public we API based on the Google search engine that shows how often a particular search-term is entered relative to the total search-volume around the world. The search trends can be filtered by country, state, category, web search/image search/news search, and time. Here we include the following terms: "influenza", "flu", "flu symptoms", "flu treatments", and "flu remedies".

We limit the search results to occur in 2012, 2013, 2014, and 2015 separately. We limit it to include the United States on the state level. Values are calculated on a scale from 0 to 100, where 100 is the location with the most popularity as a fraction of total searches in that location, a value of 50 indicates a location which is half as popular. A value of 0 indicates a location where there was not enough data for this term. A higher value means a higher proportion of all queries, not a higher absolute query count. So a tiny state where 80% of the queries are for "bananas" will get twice the score of a giant country where only 40% of the queries are for "bananas". These results are aggregated from states to U.S. Census regions in order to match up with the granularity of the MEPS data.

2.1 Data Cleaning and Feature Engineering

Survey data provides several layers of difficulty in order to transform it into a consistent and usable form. MEPS is a comprehensive survey, and in the 2015 Household data file along there are 1,832 features. The first step is to filter out many of these columns in the Household file, Medical Conditions file, and Prescribed Medicines file for each year. In the Household data we perform one-hot-encoding on the categorical variables we kept in the previous stage. For the Medical Conditions and Prescribed Medicines files we aggregate them up from the condition and prescription level to the individual level. In the Medical Conditions file we obtain counts of inpatient visits, outpatient visits, office based visits, emergency room visits, and number of prescriptions for each individual. We also create a flag indicator using ICD-9 codes if an individual had influenza. In the Prescribed medicines file we use National Drug Codes (NDC) in order to identify individuals who were prescribed Tamiflu, the only unique influenza treatment present in the dataset.

Due to the nature of conducting a survey, there are several variable values reserved for circumstances that can arise while the survey is being conducted. These values are found in Table 2.

Table 2: Reserved Value Codes in MEPS Data Files

Value	Definition
-1	Question was not asked due to skipped pattern
-2	Not asked because there was no change since previous round
-7	Question was asked and respondent refused to answer question
-8	Question was asked and respondent did not know answer
-9	Interviewer did not record the data

These values can occur in almost any variable field regardless of variable type. Some of these values are not missing at random and as such we cannot conduct imputation methods. For reserved value code -9 (Interviewer did not record the data) and -8 (Question was asked and respondent did not know answer), these values are missing at random and so we perform value imputations for this value. For binary value columns we impute the most frequent value, and for numerical columns we impute the mean. This comes out to about 8,000 imputations over all features for all years.

2.2 Data Exploration

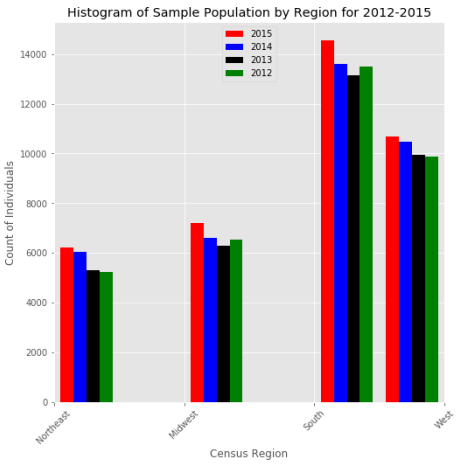
After cleaning the data and performing the above feature engineering, we end up with the count of influenza cases by year is presented in Table 3. We note the higher total frequencies in 2012 and 2013 are consistent with reporting by the Center for Disease Control (CDC) which states, "the 2012-2013 season was moderately severe, with a high percentage of outpatient visits for influenza-like illness, high rates of hospitalization, and more reported deaths attributed to pneumonia and influenza compared with recent years." (CDC, 2017) In comparison the 2014-2015 flu season according to the CDC began later than normal and as such some counts for those years are are lower.

Table 3 also contains counts of individuals with the flu and Tamiflu and without the flu but having been prescribed Tamiflu. We observe that even in severe flu seasons, Tamiflu

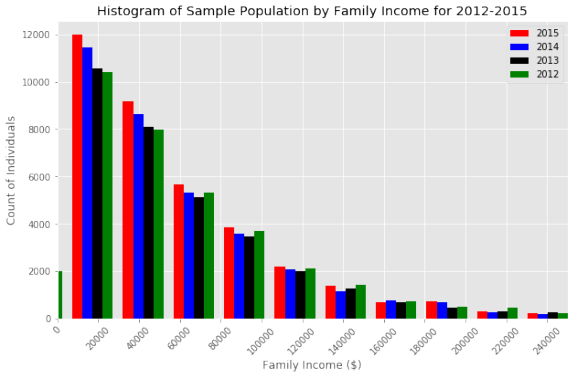
Table 3: Influenza Occurrence in Data

Year	Total Influenza Cases	Influenza with Tamiflu	Tamiflu without Influenza
2015	1,697	153	28
2014	1,803	197	38
2013	3,214	154	23
2012	3,111	86	25

prescriptions only make up a small percentage of total cases (5%-10%). We also observe less than 100 cases where an individual was prescribed Tamiflu, but did not have a flu diagnosis. This result is not surprising when dealing with survey data. If the data had been vetted by hospital and prescription records, we could justify imputation of the flu, however the survey data is based on an individuals answer to a questionnaire. We posit that these cases are mistakes either from the transcriber or from the individuals answering the questions.



(a) Histogram of Population by Region for Each Year



(b) Histogram of Family Income by Year

Figure 1: Histograms Describing the Data

Figure 1 illustrates some of the demographics of the data from the MEPS dataset for each of the years under observation. We observe consistency in demographics across the years 2012-2015. Figure 1a describes how much of each sample comes from each Census region in the country. We observe the majority of the sample comes from the South and West portions of the United States. This is fairly consistent with the distribution of population according to the U.S. Census. Figure 1b describes a histogram of family income for each of the sample years. This again follows the U.S. Census distribution with approximately 20% below $20,000 and around 50% below $100,000 in family income.

As stated above, we aggregate the Google search data proportions up from the state level to the Census region level by averaging the values for the states in each region. Table 4 shows these values for each region for each year. We can see that the Midwest has a particularly high proportion of all cases over all the years.

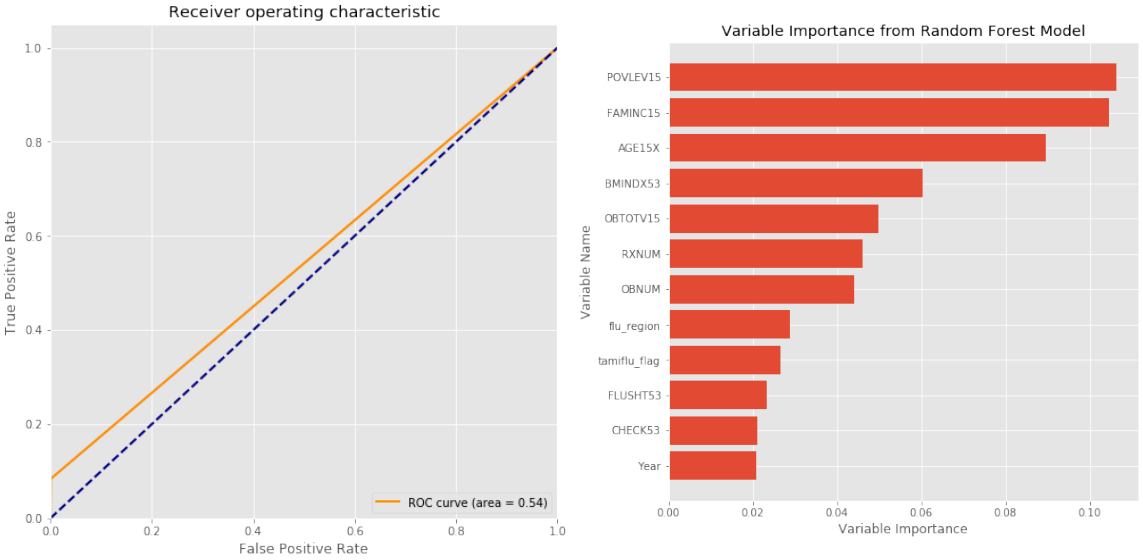
Table 4: Relative Proportion of Search Results for Flu Related Terms by Region

Year	Northeast	Midwest	South	West
2015	51%	63%	53%	50%
2014	70%	85%	80%	70%
2013	77%	82%	79%	71%
2012	50%	76%	71%	65%

3. Initial Modeling and Improving the Model

Baseline Model-Random Forest

We began by concatenating the the 2012-14 data into a training set and testing on the 2015 data. We use our created "flu-flag" from the medical conditions file as our target variable. We run a Random Forest classification model using entropy as our splitting criteria and allowing all trees to grow to their maximum depth. Figure 2a shows the results in a ROC curve with an AUC score of 0.54. Figure2b shows the variable importance for the initial Random Forest Model. Some of the top variables include: Percentage of the federal poverty level individual resides at, family income, age, BMI, number of office based visits to a health provider, number of prescription drugs used, proportion of Internet searches for flu terms in their region, whether the individual received Tamiflu, and whether the individual received a flu shot. This model does not do much better than flipping a coin so we will utilize other methods to obtain a better predictive outcome.



(a) ROC Curve with AUC

(b) Random Forest Variable Importance

Figure 2: Results from Initial Random Forest Model

Balancing the dataset

As our initial modeling show that we are barely doing better than random, we note that the data imbalance is significant in our data set. We can see from Table 5, our original data only contains 8,128 cases of flu. We investigated a few techniques for balancing the data set with synthetic data for the under-represented class and up sampling the minority class.(Chawla et al., 2002; He et al., 2008) The use of SMOTE and ADASYN for data imbalance issues has been well documented (Blagus and Lusa, 2013). We can now see from Table 5 that in both ADASYN and SMOTE methods, the classes are now completely balanced. It is important to note that the new synthetic data is generated using a nearest neighbor approach of the feature space so the synthetic data is not completely at random. (Chawla et al., 2002)

Table 5: Data Balancing By Different Methods

Method	Records Without Flu	Records With Flu
Normal Dataset	102,661	8,128
Upsampling	102,661	20,000
SMOTE	102,661	102,661
ADASYN	102,661	102,661

Choosing the right metric

Since we are using a held out test set to evaluate the performance of our models and not generating synthetic data for the test set, we still face the problem of data imbalance in test data set. In optimizing our machine learning models by grid searching for the best hyper parameters, we used F-1 score as the evaluation metric as it takes into account both the precision and recall of the model while also weighting them. This metric is more suitable for our data imbalance case because if we use a metric such as precision, the metric can be easily skewed by the majority class.

4. Results

With the balanced data set, we set out to improve the AUC ROC of our initial models. We tested a variety of machine learning models including l_2 regularized logistic regression, random forests, gradient boosted forests, support vector machines and neural networks. We have chosen not to include run the models on ADASYN dataset in the interest of time as grid searching hyper-parameters while also performing cross validation took a lot of time. We mitigated some of the time required by random searching instead of grid searching as this has been proven to be equally as good if not better. (Bergstra and Bengio, 2012) We used an instance of Amazon Web Service's C5 9X Large for models that uses the CPU while our neural network models were run on a NVIDIA GTX 1080Ti using the Keras library with a Tensorflow backend.

Model Comparison

Table 6 shows the AUC for the different models that we ran with each algorithm being run on the normal dataset, upsampled dataset and the SMOTE dataset. We can see from the table that random forests performed the best overall while SVMs did considerably worse with the highest AUC barely better than random at 0.51. It is clear that in all of the machine learning algorithms, except neural networks, either the upsampled dataset or the SMOTE dataset performs considerably better than the original dataset. We note that the highest improvement we had was for random forests with the SMOTE dataset at 0.69. This yielded an improvement gain of over 27% from our baseline model.

Table 6: Model Comparison

Area Under Curve - ROC			
	Normal Dataset	Upsampled Dataset	SMOTE Dataset
Random Forests	0.54	0.68	**0.69**
Gradient Boosted Trees	0.50	**0.58**	0.54
l_2 Logistic Regression	0.50	0.52	**0.64**
Neural Networks	**0.55**	0.50	0.54
Support Vector Machines	0.50	0.50	**0.51**

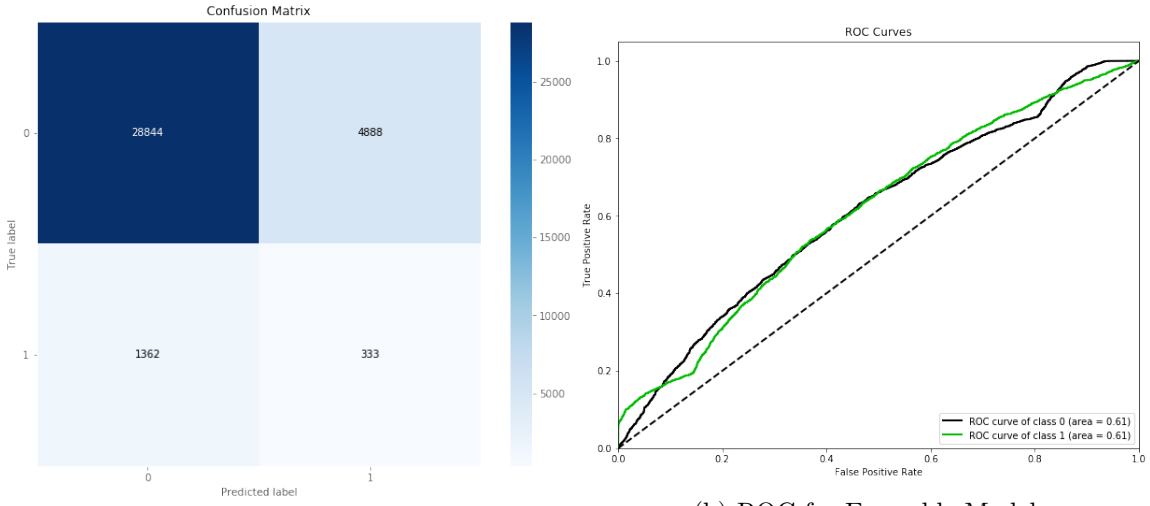
Model Ensembling

While several models, notably random forests, performed well after the balancing of the dataset, we noticed that models such as SVMs and logistic regression performed better in terms of reducing the number of false positives. However, the models did worse overall in terms of AUC score. Given these different models in prediction power of different classes, we created an ensemble of models using a weighted approach. Ensembling models is not a novel idea in machine learning and the benefit of ensembling models has been well studied. In fact, ensemble methods are currently dominate in most Kaggle competitions in terms of prediction power. (Dietterich, 2000)

5. Discussion and Conclusion

Using survey data for prediction is a precarious task because of the biases built into the survey data. For example, there is a lot more variability in the data since it relies mainly on a person's recollection. We have attempted to improve the data quality by bringing in external data sources such as Google Flu Trends data and by imputing missing responses. More importantly, most modern machine learning models assume independent and identically distributed variables (i.i.d) which is not necessarily the case in the MEPS survey data.

We had hoped that our ensemble model would perform best overall but we can clearly see from Figure 3 that even though we had lower false positives, we had a much higher false negatives than any individual models. We can also see that the ensemble model did not do the best in terms of AUC as we can see in Figure 3b. However, we do note that this ensemble model is only outperformed by upsampled, SMOTE random forests and l_2 regularized logistic regression. This could be mainly attributed to the small number



(a) Ensemble Model Confusion Matrix

(b) ROC for Ensemble Model

Figure 3: Ensemble Model Metrics

of weights that were used to calibrate the ensemble model of the different base learners. (Dietterich, 2000)

However, we are pleased with the improvement of over 37% from our baseline model in terms of AUC. We achieved this through careful imputation and using synthetic data generation methods such as SMOTE to balance the heavily imbalanced dataset. Possible avenues of future research could be into the use of modern machine learning for survey data without violating the i.i.d assumptions. We could investigate further into the use of text mining methods to improve the prediction of our algorithms using a base learner such as neural networks that could combine both the text analysis and the survey data analysis.

References

AHRQ. Medical expenditure panel survey. *Medical Expenditure Panel Survey*. URL `https://meps.ahrq.gov/mepsweb/`.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13(1):281–305, February 2012. ISSN 1532-4435.

Rok Blagus and Lara Lusa. Smote for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 14(1):106, Mar 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-106. URL `https://doi.org/10.1186/1471-2105-14-106`.

CDC. Influenza (flu). Oct 2017. URL `https://www.cdc.gov/flu/about/index.html`.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, Jun 2002.

Thomas G. Dietterich. Ensemble methods in machine learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, MCS '00, pages 1–15, London, UK, UK, 2000. Springer-Verlag. ISBN 3-540-67704-6.

Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):10121014, Feb 2009. doi: 10.1038/nature07634.

Haibo He, Yang Bai, Edwardo A. Garcia, and Chengchao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328, 2008.

D. Lazer, R. Kennedy, G. King, and A. Vespignani. The parable of google flu: Traps in big data analysis. *Science*, 343(6176):1203–1205, Mar 2014. doi: 10.1126/science.1248506.

Sarah F. Mcgough, John S. Brownstein, Jared B. Hawkins, and Mauricio Santillana. Forecasting zika incidence in the 2016 latin america outbreak combining traditional disease surveillance with search, social media, and news report data. *PLOS Neglected Tropical Diseases*, 11(1), 2017. doi: 10.1371/journal.pntd.0005295.

Melissa A Rolfes, Ivo M Foppa, Shikha Garg, Brendan Flannery, Lynnette Brammer, James A Singleton, Erin Burns, Daniel Jernigan, Carrie Reed, Sonja J. Olsen, and et al. Estimated influenza illnesses, medical visits, hospitalizations, and deaths averted by vaccination in the united states. *Centers for Disease Control and Prevention*, Apr 2017. URL `https://www.cdc.gov/flu/about/disease/2015-16.htm`.