Final Project Report for Stats 170AB, Winter/Spring 2019

Project Title: Predict Frequencies of Keywords Related to Health and Activities in Tweets

Lizhi Feng 80189236 lizhif1@uci.edu
Xinyi Wei 75707053 weix7@uci.edu

Github: https://github.com/XinyiWei/Final-Project-STAT170

1. Introduction and Problem Statement

Air pollution is a significant risk factor for respiratory illnesses and diseases. The risk is increased when people are active because exercise and physical activity cause people to breathe faster and more deeply and to take more particles into their lungs. So we did research on the relationship between daily air quality and public participation in outdoor activities in the city level. We used the frequencies of some keywords from tweet status updates for reflecting related public activity response and perceptions. The goal was to determine which features effect public participation in outdoor activities and perceptions of health.

Evidence suggested a weak relationship between the increase of air pollutions and the decrease odds of physical inactivity in California, while population size and city area are the most effective features. After normalizing tweet counts by population, we found that the increase in physical activity and health concerns are associated with increased city level median family income and outdoor-activity-friendly level among the normal weight.

2. Related Work

Recent studies have shown the utility of air quality data sources for a wide range of public health goals, as “The Association of Ambient Air Pollution and Physical Inactivity in the United States” (Jennifer D. Roberts, Jameson D. Voss, Brandon Knight) demonstrates. They perform logistic regression models as evidence that air pollution should be investigated as an environmental determinant of inactivity. The findings of their research also emphasized the phenomenon that there is a complex interplay among many risk factors, behavioral and demographic variables, which are associated with physical activity.

3. Data Sets

a. Tweet status updates dataset

The Tweet Status Updates dataset includes status posts made by users in California from January 2016 to march 2019 and filtered by the following keywords.

Keywords: **air pollution, pollutant, fume health, asthma, respiratory, pneumonia outdoor, activity, activities, hiking, biking, jogging, running, climbing, camping, fitness**

b. Weather dataset (65535 Rows x 5 Columns)

DATE	LATITUDE	LONGITUDE	ELEVATION	PRCP	TMAX
1/1/16	37.8947	-122.2965	16.8	0	
1/2/16	37.8947	-122.2965	16.8	0	
1/4/16	37.8947	-122.2965	16.8	0.05	
1/5/16	37.8947	-122.2965	16.8	0.7	
1/6/16	37.8947	-122.2965	16.8	0.96	
1/7/16	37.8947	-122.2965	16.8	0.72	
1/8/16	37.8947	-122.2965	16.8	0	
1/9/16	37.8947	-122.2965	16.8	0.09	
1/10/16	37.8947	-122.2965	16.8	0.02	

DATE	CITY	ELEVATION	PRCP	TMAX
1/1/16	Albany	16.8	0	
1/2/16	Albany	16.8	0	
1/4/16	Albany	16.8	0.05	
1/5/16	Albany	16.8	0.7	
1/6/16	Albany	16.8	0.96	
1/7/16	Albany	16.8	0.72	
1/8/16	Albany	16.8	0	
1/9/16	Albany	16.8	0.09	
1/10/16	Albany	16.8	0.02	

The dataset includes observations of **elevation**, daily **total precipitation** and **max temperature** from over 1000 stations in California.

References: National Centers for Environmental Information
https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc:C00861/html

c. Air Quality dataset (162193 Rows x 4 Columns)

DATA	LATITUDE	LONGITUDE	COUNTY	AQI_VALUE
1/1/16	37.687526	-121.784217	Alameda	41
1/2/16	37.687526	-121.784217	Alameda	53
1/3/16	37.687526	-121.784217	Alameda	72
1/4/16	37.687526	-121.784217	Alameda	61
1/5/16	37.687526	-121.784217	Alameda	21
1/6/16	37.687526	-121.784217	Alameda	17
1/7/16	37.687526	-121.784217	Alameda	40
1/8/16	37.687526	-121.784217	Alameda	44
1/9/16	37.687526	-121.784217	Alameda	40
1/10/16	37.687526	-121.784217	Alameda	44

DATA	CITY	COUNTY	AQI_VALUE
1/1/16	Livermore	Alameda	41
1/2/16	Livermore	Alameda	53
1/3/16	Livermore	Alameda	72
1/4/16	Livermore	Alameda	61
1/5/16	Livermore	Alameda	21
1/6/16	Livermore	Alameda	17
1/7/16	Livermore	Alameda	40
1/8/16	Livermore	Alameda	44
1/9/16	Livermore	Alameda	40
1/10/16	Livermore	Alameda	44

The **AQI** is an index for reporting daily air quality. It indicates how clean or polluted the air is, and what associated health effects might be a concern for people.The higher the AQI value, the greater the level of air pollution and the greater the health concern. We transferred the GPS coordinates into city name.

References: the United States Environment Protection Agency.
https://www.epa.gov/outdoor-air-quality-data

d. Walk Score & Bike Score (372 ROWS x 3 Column)

City	Walk Score	Bike Score
Adelanto	9	28
Agoura Hills	36	--
Alameda	65	73
Albany	80	--
Alhambra	69	47
Aliso Viejo	37	34
Altadena	49	38
American Canyon	34	--
Anaheim	54	49
Antelope	34	43

The **walk score** and **bike score** reports how friendly a city is to walking and biking. There are 372 largest California cities in this table. Cities with higher scores are more suitable and safer for walking and biking.

References: https://www.walkscore.com/CA/

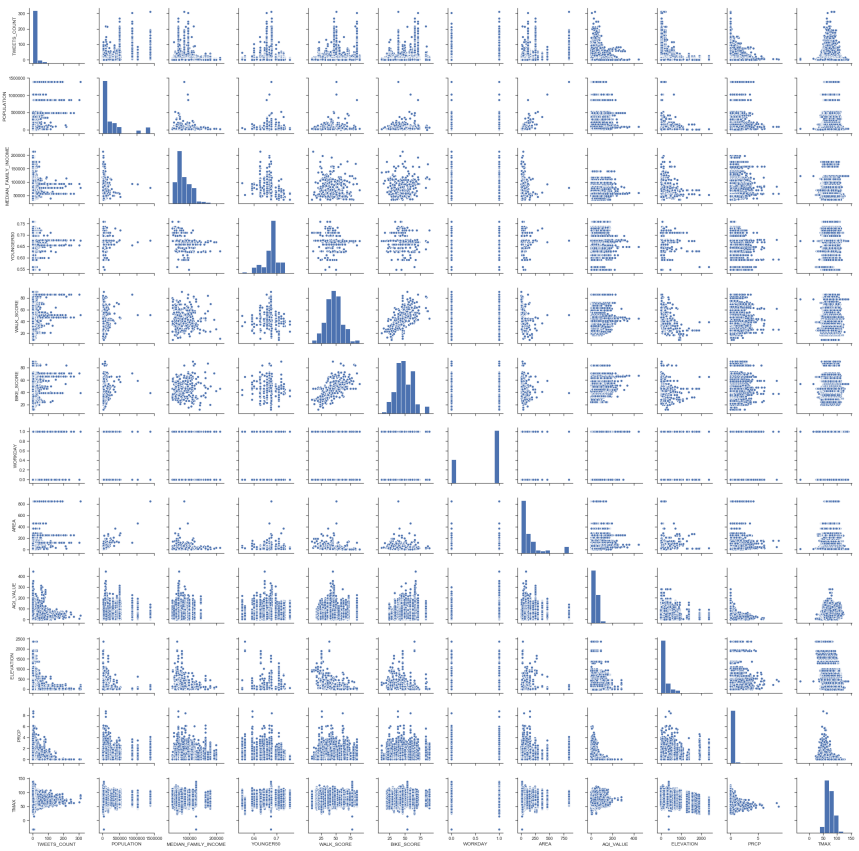
e. Others

CITY	SEASON	DATE	POPULATION	AREA	WORKDAY	ELEVATION	INCOME	UNDER50	WALK_SCORE	BIKE_SCORE	AQI	PRCP	TMAX	TWEETS_COUNT
Alameda	winter	1/1/16	78246	27.5	0	155.0450813	93427	0.676552834	65	73	50.27169689	0	57.35667752	2
Berkeley	winter	1/1/16	120179	27.1	0	155.0450813	114720	0.676552834	81	85	50.27169689	0	57.35667752	6
Dublin	winter	1/1/16	57022	38.6	0	155.0450813	128737	0.676552834	34	56	50.27169689	0	57.35667752	4
Fremont	winter	1/1/16	230964	200.6	0	155.0450813	112347	0.676552834	44	49	50.27169689	0	57.35667752	1
Hayward	winter	1/1/16	156917	117.4	0	155.0450813	69415	0.676552834	53	52	50.27169689	0	57.35667752	1
Livermore	winter	1/1/16	88232	65.2	0	155.0450813	111950	0.676552834	37	64	50.27169689	0	57.35667752	1
Newark	winter	1/1/16	45554	35.9	0	155.0450813	91103	0.676552834	47	45	50.27169689	0	57.35667752	4
Oakland	winter	1/1/16	417442	144.5	0	155.0450813	60472	0.676552834	72	65	50.27169689	0	57.35667752	8
San Leandro	winter	1/1/16	89910	34.6	0	155.0450813	76781	0.676552834	62	55	50.27169689	0	57.35667752	1

To build a reliable predictor, we considered some demographic characteristics which may effect the results. **Population size**, **median family income**, **percentage under 50**. In addition, since weather and activity are seasonal, we added season indicator and workday indicator based on the date.

After merging data by city and date, we got the full data (180458 Rows x 15 Columns)

The Cross-correlation Plot



4. Technical Approach

a. Data Processing

To begin, the tweets status updates dataset was transformed from JSON file into Apache Parquet, an open-source column-oriented data storage format, then imported to Pandas data frame. Its compression is efficient and saves us storage space, and the processing is fast. We counted the daily total number of tweets from each city and made a new dataset.

For the weather and air quality data, latitudes and longitudes are translated into city names with the help of pygeocoder method from Geocoder. From here, the datasets can be merged by the city name and recorded date. Each row in the full dataset represented one day for one city. We removed duplicates so that each city is represented only once in one day, no matter the number of observation sites it has.

Different approaches were used to handle missing data. For time serious problem, we considered the effect of season and trend. Missing datas in weather are replaced by daily average, while those in AQI are replaced by local average. For general problems like incomes and age, we simply used the mean value.

b. Data Analysis

After preparing the data, it was time to build models. We tried linear regression and Poisson regression first. Even the R^2 scores were pretty high (around 0.70), there was a large gap in both prediction distributions. Since the tweets count is a numeric variable, to use logistical regression, we split it into two groups. Then we got a new target variable — 'large count' In the data frame. 1 means relatively more people in the city mentioned keywords that day. The models will be evaluated using the percentage accuracy of the classifiers. To get this accuracy, a train-test split where we train the model on 75% of the data and test on the remaining 25%. To improve accuracy, we applied K-fold cross validation and divided the data into 10 groups. In the first iteration, the first fold is used to test the model and the rest 9 serve as the training set. This process is repeated until all 10 folds have been used as the testing set. Finally, we focused on random forests. A random forest with 20 decision trees was the best model we could found for our data set. We got the importance of each variable according to our random forest model as well.

5. Software

Software	Usage
Matplotlib.pyplot	Draw Plots
Pandas	Manipulating Dataframe
patsy	Dmatrices
Pyspark	Loading Json Twitter Model
Sklearn	LogisticRegression
Sklearn	Metrics (calculate MSR)
Sklearn	Linear.linear_regression
Sklearn.ensemble	RandomForestClassifier
Sklearn.linear_model	Initiate linear models
Sklearn.metrics	Model Accuracy_Score
Sklearn.model_selection	Train_test_split
Statsmodels.api	GLM Model
Yellowbrick.regressor	ResidualsPlot (Linear)
Jupyter	Run all codes above
Tableau	Draw Plots

In this project, we used Jupyter, tableau, Matplotlib, Pandas, Pasty, Pyspark, Sklearn, Statsmodels, and Yellowbrick. These are all useful tools, but I think Sklearn and Yellowbrick is the most useful tool in this project. Sklearn has many built-in algorithms that like logistics model, linear regression model, random forest model, and GLM model. Because python has those models, we do not need to calculate models by formulas. Another built-in module that surprised me is Yellowbrick because it can build linear regression and plot the residuals verse predictions immediately. If we do not have those model functions, it will hard for us to get any prediction.

6. Experiments and Evaluation

a. The First Experiments

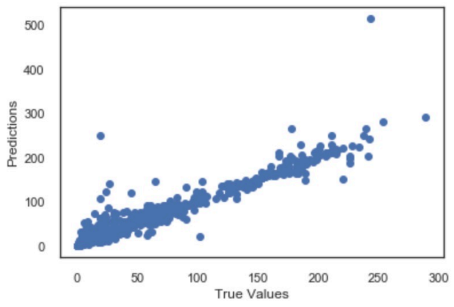
According to the pair plot between counts and the other twelve variables. We used four models: linear regression model, logistic model, general linear model (Poisson Distribution) and Random forests model. At this part, we use the normalized population data, because the population values are too large to affect the results. The R^2 is the measure of how close the data are to the fitted values. (Formula:$R^2 = 1 - \frac{SS_{RES}}{SS_{tot}}$). MSE is squared of the R^2, and model accuracy is how many percentages of perditions same as true values.

We got the following results as below.

	R^2	MSE	Model Accuracy
BaseLine	0.136	0.369	0.864
Linear Regression	68.79	8.29	0.64
Logistic Regression	0.0996	0.32	0.9
GLM(Poisson)	52.36	7.24	0.72
Random Forest	3.5	1.87	0.95

For determine the baseline, the purpose of the project is to evaluate if twelve variables affect the outdoor activities. Then, adding the column counts people mention keywords of activities per one hundred thousand population. After getting the 'Count_by_population' variable, we get the 50% of 'Count_by_population' is 2.773, so we use counts less than 2.773 as an indicator that twitter counts are not related to other twelve predictor variables. After fitting logistic regression with binary response variable (if less than 2.77 counts), we get the result that R^2 is 0.136, MSE is 0.369, and Model Accuracy is 0.864.

The first experiment that we fit the model with twelve variables: 'COUNT', 'INCOME', 'percentage', 'Walk_Score', 'Bike_Score', 'WORKDAY', 'AREA', 'AQI_VALUE', 'ELEVATION', 'PRCP', 'TMAX', 'season', 'Count_by_population'. Among those variables, we transfer season variables to numerical variables. For these four models, we split the training and testing data as 80:20. In the meanwhile, for the random forest model, we set n_estimator =20. We also use binary variables (if less than 6) as the response variable for the logistic regression model.




```
K-fold Logistic Regression Scores

[0.8155824005319738,
 0.7806716169788319,
 0.7848276626399202,
 0.8253352543499944,
 0.8389116701762164,
 0.8440651667959659,
 0.8558129225313089,
 0.863958772027042,
 0.8624549736769188,
 0.8665558326406206]


 K-fold Logistic Regression Error Rates

 [0.18441759946802616,
  0.21932838302116814,
  0.2151723373600798,
  0.17466474565000553,
  0.16108832982378365,
  0.15593483320403415,
  0.1441870774686911,
  0.136041227972958,
  0.1375450263230812,
  0.13344416735937933]
```

b. The Second Experiments

In the second experiments, we split the train and test data to 60:40. The result we got is below.

	R^2		MSE		Model Accuracy	
Split	80 :20.	60:40.	80 :20.	60:40.	80 :20.	60:40.
Linear Regression	68.79	72.12.	8.29	8.49.	0.64	0.65
Logistic Regression	0.0996	0.056.	0.32	0.238	0.9	0.943
GLM(Poisson)	52.36	54.42.	7.24	7.37	0.72	0.72
Random Forest	3.5	6.78.	1.87	2.61	0.95	0.945

For the linear regression model, GLM and random forest models, the model accuracy increase. However, for logistic models, the model accuracy decreases. Because we use binary response variables here for logistic models, the less training data will definitely affect the model accuracy, R^2, and MSE.

c. Third Experiment

Changing the size of training and testing data does not change a lot to results. Therefore, we think that changing some variables that like season (winter:0, spring:1, summer:2, fall:3) or WORKDAY (holiday:0, workday:1) as dummy variables may help. Therefore, we did the third experiment. The result is below.

	R^2		MSE		Model Accuracy	
Dummy or not	Not Dummy	Dummy	Not Dummy	Dummy	Not Dummy	Dummy
Linear Regression	68.79	64.38.	8.29	8.023	0.64	0.644
Logistic Regression	0.0996	0.054.	0.32	0.231	0.9	0.946
GLM(Poisson)	52.36	72.59.	7.24	8.52	0.72	0.64
Random Forest	3.5	4.95	1.87	2.22	0.95	0.953

From the experiment, we could see the R^2, MSE and model accuracy do not change a lot due to the change of dummy variables. Linear regression, logistic regression, and random forest models increase a little model accuracy. However, the GLM (Poisson) model does not change a lot for its MSE and accuracy.

After three experiments, we could know that our model is not sensitive by its number of train and test datasets. In the meanwhile, it is not sensitive for the season and workday dummy variable as well. Among the four models above, the random forest model gets the best score. The MSE for this random forest model is 1.87 because the average of amounts of tweets is about 5. However, there are some cities like San Diego and Los Angeles, and their average tweets count are about 150. Therefore, those big cities will affect our models and generate MSE.

7. Notebook Description

The notebook that we have basically contains two major components: cleaning data and building models. In the cleaning data part, we merge all nine data sources to a table with180458 rows and 16 columns. Then we also have the part that normalizes the population

data by calculating ‘Count_by_population (per ten thousand people)’ by using the formula: 100000*counts/population. And in the building model part, we have four models linear, regression model, logistic model, GLM model and random forest model to train. In the meanwhile, there are also parts that divide train: test to 60:40, train: test=80:20, and dummies for season and WORKDAY variables for four models. Another major part in building model notebook is that we plot predictions verse true values plots for all models, and there is a residuals plot for the linear regression model as well.

8. Members Participation

Our project basically has three major steps: finding data, cleaning and merging data, and building models. For finding data part, we contact Professor Chen, Li to get the Twitter data together. Then we plan to look for four data sources online per person. After gathering all datasets, we need to clean and merging data. Lizhi did most part forfending and cleaning data such as filtering Nan data and normalizing data. Xinyi did most part to merge data such as connecting all database appropriately and group them well. In building models part, Lizhi did logistics regression and linear regression model, and Xinyi did random forest and GLM models.

Participation		
Task	Member	Percentage
Finding Twitter Data	Lizhi Feng, Xinyi Wei	50% : 50%
Finding Population Data	Lizhi Feng	100%
Finding Income Data	Lizhi Feng	100%
Finding Area Data	Lizhi Feng	100%
Finding Air Quality Data	Lizhi Feng	100%
Finding Gender Data	Xinyi Wei	100%
Finding Walk/Bike Score Data	Xinyi Wei	100%
Finding Holiday Data	Xinyi Wei	100%
Finding Weather Data	Xinyi Wei	100%
Cleaning Data	Lizhi Feng, Xinyi Wei	70% : 30%
Merging Data	Xinyi Wei, Lizhi Feng	70% : 30%
Building Models (Logistic, linear regression)	Lizhi Feng	100%
Building Models (Random Forest, GLM)	Xinyi Wei	100%

9. Discussion and Conclusion

In the project, we use the linear regression algorithm, GLM algorithm, random forest algorithm, and logistic regression algorithm. In the meanwhile, we use methods like df.merge, df.groupby and train_test_split. Among the four models that we use, linear regression is the easiest to understand and explain, but it does not fit well in the non-linear model. The logistic model can give us a good interpretation for binary results, but it does not seem to be very flexible. We think the random forest model will not be affected a lot by dummy variables, but the model is hard to interpret.

Cleaning and merging data spend more time than we expect. We think that two weeks is enough to clean and merge data, but we spend like 6 weeks to merge all databases. Originally, we think training models will consume the most time, but it actually spends only two weeks.

During finishing the project, I also learned some other methods that named fillna and geo_coder. 'Fillna' method could fill nan data with a specific value. The most interesting method is geo_coder which could transform longitude and latitude to a specific city name.

If we were in charge of a research lab, it would be better if we could divide the keywords into groups to see the association between air quality and each specific topic. We think the model and result would get more accurate. Meanwhile, we may consider predicting which word has the most possibility to occur on Twitter. If we have more time, we may extend our database not only in the California region but also in all the USA, since the air quality and climate condition in California are good enough for outdoor activities for the most of times. A larger dataset from the whole country would help us to investigate the relationship better.