A Machine Learning Framework to predict the prevalence of Skin Cancer across US from Digital Records of online behavior (January 2018)

Sibghat Ullah[1], Siddhant Tandon[2]

[1,2]Sapienza University of Rome, Department of Information Engineering, Automation and Management, 00185 Rome, Italy

I. Abstract

UNDERSTANDING the behavior of diseases with the help of online records is becoming popular among the Digital Epidemiologists. This document aims to predict the behavior of Skin Cancer across the United States with the help of online record from search engines. More precisely, we study the correlation of online record with our ground truth. Later, we try to predict the behavior of Skin Cancer with this online record. We also investigate the importance of various factors including Poverty, Income and Health Insurance. We make a brief comparison of our Machine Learning models in terms of their prediction performance. We eventually conclude that this online digital data from Search Engines is not sufficient to build reliable Machine Learning model(s) to predict Skin Cancer.

II. Introduction

The aim of this project is to find correlation between online record of Google and the ground truth data on Skin Cancer across the United States. We also build the Machine Learning models based on the data from Google and compare these Models to predict the behavior of Skin Cancer across the United States.

The Ground Truth Data is acquired from US Center for Disease Control and prevention. The data set is divided in two categories namely 'Age Adjusted' and 'Crude'. Furthermore, this data is related to six years from 2011 to 2016. Please note that the data set used in this project are aggregated for the USA states and Washington DC. This ground truth data is our benchmark data that we'll use to compare our correlation and Machine Learning models. The online digital record is acquired from Google trend data based on three search terms 'Skin Cancer', 'Melanoma' and 'Carcinoma' respectively. The reason to introduce 'Melanoma' and 'Carcinoma' is that these terms are scientific synonyms for 'Skin Cancer' and we might find them useful in predicting the behavior of 'Skin Cancer'. We also use other data sets related to Income, Poverty and Health Insurance from US Census Bureau and Wikipedia to see the impact of these factors in predicting our disease of interest i.e. 'Skin Cancer'. The detailed description of these data sets is included in the tasks which use them.

III. Correlation

A The first task is to find the correlation between our Ground Truth Data and online data from Google to find if it's reasonable to build Advance Machine Learning / Data Mining models. We choose 'Pearson Correlation' as our guiding Statistics in this task. The software which performs this task is written in Python and is available in folder 'Source Code'. Remember that we find the correlation for 'Skin Cancer', 'Melanoma' and 'Carcinoma' data respectively with the ground truth data for six years i.e. 2011,2012,2013,2014,2015 and 2016. Furthermore, we perform this activity two times i.e. when the ground truth data is 'Age Adjusted' and when it's 'Crude'.

The Correlation achieved in this activity is discussed in Table 1 and Table 2. To have an overall understanding of the correlation, please see figures 1-8.

B In this task, we perform the correlation between the Ground truth data and the US census bureau data on income and poverty. We consider three different cases in this task.

Case B.1 In this case, we consider three-year Average poverty rate (2014-2016) for the entire US and find it's comparison with three-year Average (2014-2016) Skin cancer rate for the entire US. We find out that the Average poverty rate has a correlation of 0.13 for 'Age Adjusted' Skin cancer rate while it has a 0.02 correlation for 'Crude' Skin cancer rate. We describe the documentation for this process in Table 3.

Case B.2 In this case, we consider four different US income levels. These income levels are 'Fulltime Male USA', 'Fulltime Female USA', 'Seasonal Male USA', and 'Seasonal Female USA'. We find the correlation between the ground truth data and the income

data. Remember that the correlation here is aggregated temporally (2011-2016) and spatially (all US States and Washington DC). As always, we perform this activity for both categories of ground truth data i.e. 'Age Adjusted' and 'Crude'. We find out that all the values of correlation for both ground truth data types are negative. The highest correlation for 'Age Adjusted' Skin cancer is with the Income category 'Seasonal Male USA'. Similarly, the highest correlation for 'Crude' Skin cancer is with the same income category. Another important thing is that the correlation of 'Age Adjusted' Skin cancer is acting as an upper bound for 'Crude' Skin cancer. The process is documented in Table 4 and Figure 9.

Case B.3 In this case, we consider the median annual income data for all US states and Washington DC from 2011 to 2015 and find the correlation of ground truth data spatially. As always, ground truth data will involve both types 'Age Adjusted' and 'Crude'. We find out that the values of correlation vary between -0.88 and 0.94. The correlation of median income average data is generally lower for 'Age Adjusted' Skin cancer but not always. Considering the ground truth data type 'Age Adjusted', its highest correlation with income data is for the state 'North Dakota' while its lowest correlation with income data is for the state 'Texas'. On the other hand, Considering the ground truth data type 'Crude' the highest correlation between the ground truth data and income data comes for the state 'Kentucky' while the lowest correlation is found for 'Pennsylvania'. If we don't differentiate between the types of ground truth data and look globally, the highest correlation is for the state 'Kentucky' while the lowest correlation is for 'Texas'. The average correlation for 'Age Adjusted' is 0.21 while for 'Crude' data type, it is 0.5. For reference, please see Table 5 and Figure 10 to understand completely.

IV. Prediction

A This task aims to build Machine learning model(s) to predict our ground truth data with the help of online record of search terms from Google. Thus, our independent variables are the data acquired from Google trend data for three keywords namely 'Skin Cancer', 'Melanoma' and 'Carcinoma'. The reason to add 'Melanoma' and 'Carcinoma' was that these terms are scientific synonyms for 'Skin Cancer'. Thus, they might be helpful in predicting the behavior of 'Skin Cancer' across the US. We also tried to extract the data for other search terms like 'Skin Cancer Cure' or 'Skin Cancer Therapy' but Google denied our request generating an error that the Data is insufficient for all the US states and thus can't be provided. To this end, we have 18 features in total corresponding to three different search terms 'Skin Cancer', 'Melanoma' and 'Carcinoma' and six different years 2011,2012,2013,2014,2015 and 2016. Our ground truth data (to be predicted) has 12 variables in total corresponding to two distinct types 'Age Adjusted' and 'Crude' and six different years same as above. We perform LASSO regression with Spatial and temporal Cross validation. We have different choice of parameters to be selected for an optimal and reliable model. We perform Grid Search Cross Validation where cross Validation is with respect to US States. In other words, we fix a specific year i.e. 2011 etc. and then we use the features of online data i.e. 'Skin Cancer', 'Melanoma' and 'Carcinoma' for that specific year and try to predict the Ground truth data i.e. 'Age Adjusted Skin Cancer', 'Crude Skin Cancer' for that specific year and we repeat this process six times as we have six different years. Remember that for each year, we have a different machine learning model which is selected to minimize the average Cross validated error. We mention here the 'Mean Absolute Error' in detail and denote it as 'MAE'.

The average 'Mean Absolute Error' for all the six year is '.65. This means that if we pick a random US state for any of the six years, we'll predict the True value of skin cancer in that state as True Value ± 0.65. Furthermore, the average 'Mean Absolute Error' for the ground truth 'Age Adjusted Skin Cancer' is 0.63 while for the 'Crude Skin Cancer', this error is 0.67. Also, the maximum 'MAE' for 'Age Adjusted Skin Cancer' is 0.70 while for 'Crude Skin Cancer', the corresponding value is 0.81. We also mention here the minimum values of 'MAE' which are 0.52 and 0.48 for 'Age Adjusted Skin Cancer' and 'Crude Skin Cancer' respectively. We provide Table 6 and figure 11 for more details.

B The second task aims to include the income features for the US states and see the impact on the predictive performance of the model. We extract the data from the US Census bureau related to the median annual income of all US states and Washington DC. However, we were able to receive data for five years only which are 2011,2012,2013,2014 and 2015. This means that we will not be able to include the year '2016' in our analysis. Thus, from now on we'll only consider the data related to above mentioned five years. As above, we perform Lasso regression with the same setup which means that the model will choose the best parameters based on Spatial cross validation and we'll repeat this process five times i.e. for five distinct years. However, this time we have Income features i.e. Median income as

well and we're interested in finding out the impact of those features. We mention the details of 'MAE' in Table 7 while Figure 12 explains the process in greater detail.

C The third task aims to include the features related to Health Insurance in our Machine learning model (which already includes the online google trend data for three different search terms and the median annual income for all the US states) and understand the prediction performance associated in the process. We extract the data from US census bureau and Wikipedia manually and save them in one table. This data is the percentage of people who are uninsured in all the US states and the capital. We have this data for years 2011,2012,2013,2014 and 2015. Thus, this is like process '*B*' and the only difference is that now we also have health insurance features in our model as well. We perform the regression again and mention the details of 'MAE' in Table 8 and figure 13.

V. TABLES AND FIGURES

A. Tables

Table 1

Ground Truth Data	**Google Trend Data**	**Max Correlation**	**Min Correlation**	**Average Correlation**
Age Adjusted	Skin Cancer	0.49	0.39	0.43
Age Adjusted	Melanoma	0.33	0.12	0.23
Age Adjusted	Carcinoma	0.39	0.17	0.27
Crude	Skin Cancer	0.11	-0.07	-0.014
Crude	Melanoma	-0.04	-0.15	-0.08
Crude	Carcinoma	-0.15	-0.24	-0.18

Table 1 explains the Pearson correlation coefficient between Ground Truth Data and online Google trend data for three keywords. Please note that here, we are considering the correlation coefficient which is aggregated for all US states.

Table 2

Ground Truth Data	**Google Trend Data**	**Max Correlation**	**Min Correlation**	**Average Correlation**
Age Adjusted	Skin Cancer	0.91	-0.13	0.46
Age Adjusted	Melanoma	0.73	-0.91	-0.22
Age Adjusted	Carcinoma	0.55	-0.83	-0.28
Crude	Skin Cancer	0.86	-0.6	0.25
Crude	Melanoma	0.9	-0.82	-0.13
Crude	Carcinoma	0.87	-0.8	0.0

Table 2 explains the Pearson correlation coefficient between Ground Truth Data and online Google trend data for three keywords. Please note that here, we are considering the correlation coefficient which is aggregated for all the years (2011-2016).

Table 3

Data Type	**Average Correlation**	**Years**
Age Adjusted	0.13	2014-16
Crude	0.02	2014-16

Table 3 explains the average correlation between Ground Truth Data and Poverty Ratio in all the US states and Capital. The correlation is aggregated Spatially and Temporally.

Table 4

Data Type	**US Income Category**	**Correlation**
Age Adjusted	Male Full Time	-0.54
Age Adjusted	Female Full Time	-0.66
Age Adjusted	Male Seasonal	-0.41
Age Adjusted	Female Seasonal	-0.61
Crude	Male Full Time	-0.91
Crude	Female Full Time	-0.95
Crude	Male Seasonal	-0.89
Crude	Female Seasonal	-0.97

Table 4 explains the Average Correlation between the four distinct US Income categories and the Ground Truth. Please note that the data regarding the categories is aggregated Spatially and Temporally.

Table 5

Ground Truth Data	Annual Median Income	Average Correlation	Max Correlation	Min Correlation
Age Adjusted	All US States (2011-16)	0.21	0.93	-0.88
Crude	All US States (2011-16)	0.5	0.94	-0.73

Table 5 explains the Correlation between the Ground Truth Data (Skin Cancer ratio) and Income Data (median annual income). This is aggregated over the years (2011-2016) and is represented State wise.

Table 6

Ground Truth Data	Average MAE	Max MAE	Min MAE
Age Adjusted	0.63	0.7	0.52
Crude	0.67	0.81	0.48

Table 6 explains the Prediction performance of Machine Learning model which is based on the online Google trend data in terms of 'Mean Absolute Error' aka 'MAE'.

Table 7

Ground Truth Data	Average MAE	Max MAE	Min MAE
Age Adjusted	0.63	0.72	0.54
Crude	0.72	0.80	0.58

Table 7 explains the Prediction performance of Machine Learning Model which is based on Google trend data and median annual income for US states and capital in terms of 'MAE'

Table 8

Ground Truth Data	Average MAE	Max MAE	Min MAE
Age Adjusted	0.63	0.72	0.54
Crude	0.73	0.83	0.58

Table 8 explains the Prediction performance of Machine Learning Model which is based on Google trend data, median annual income for US states and capital and percentage of uninsured people in USA in terms of 'MAE'.

B. Figures

Figure 1

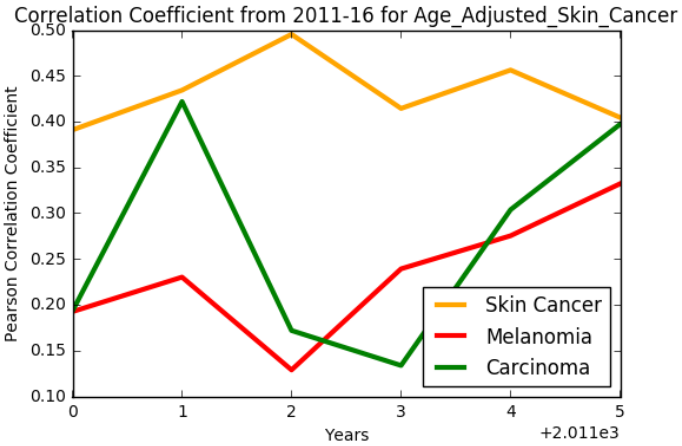

Figure 1 explains the Correlation between our Ground Truth (Age Adjusted) and Google Trend data for three key words namely 'Skin Cancer', 'Melanoma' and 'Carcinoma respectively'. The correlation is from year 2011-2016.

Figure 2

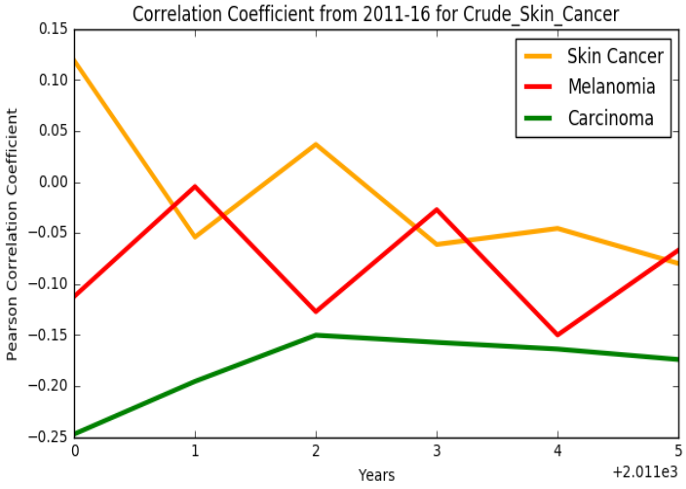

Figure 2 explains the Correlation between our Ground Truth (Crude) and Google Trend data for three key words namely 'Skin Cancer', 'Melanoma' and 'Carcinoma respectively'. The correlation is from year 2011-2016.

Figure 3

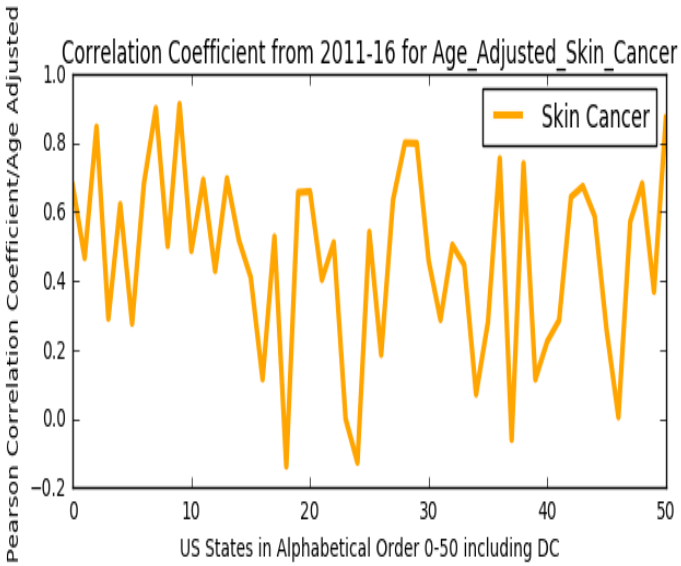

Figure 3 explains the Correlation between our Ground Truth (Age Adjusted) and Google trend data for search term 'Skin Cancer'. This correlation is presented here as State wise (all US States and Washington DC) and is aggregated temporally (2011-2016). Please also note that US states are sorted alphabetically and their corresponding rank in this sorted order is presented here instead of the name.

Figure 4

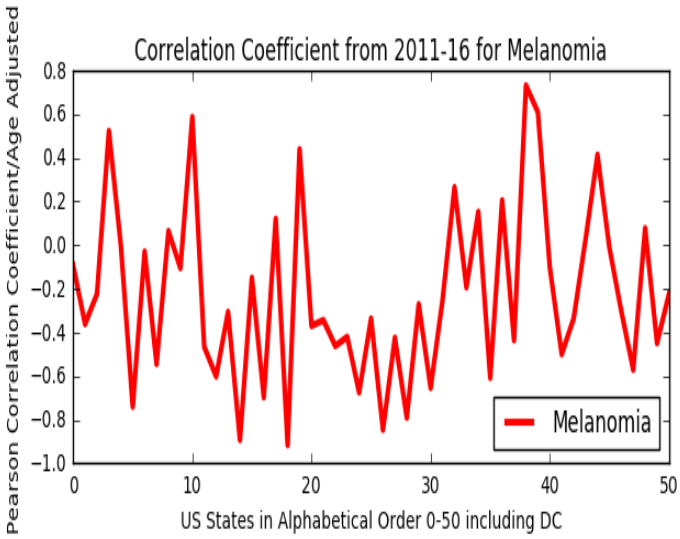

Figure 4 explains the Correlation between our Ground Truth (Age Adjusted) and Google trend data for search term 'Melanoma'. This correlation is presented here as State wise (all US States and Washington DC) and is aggregated temporally (2011-2016). Please also note that US states are sorted alphabetically and their corresponding rank in this sorted order is presented here instead of the name.

Figure 5

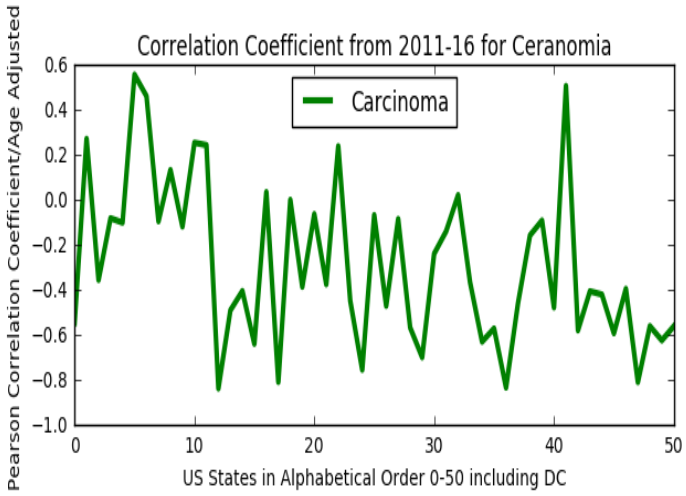

Figure 5 explains the Correlation between our Ground Truth (Age Adjusted) and Google trend data for search term 'Carcinoma'. This correlation is presented here as State wise (all US States and Washington DC) and is aggregated temporally (2011-2016). Please also note that US states are sorted alphabetically and their corresponding rank in this sorted order is presented here instead of the name.

Figure 6

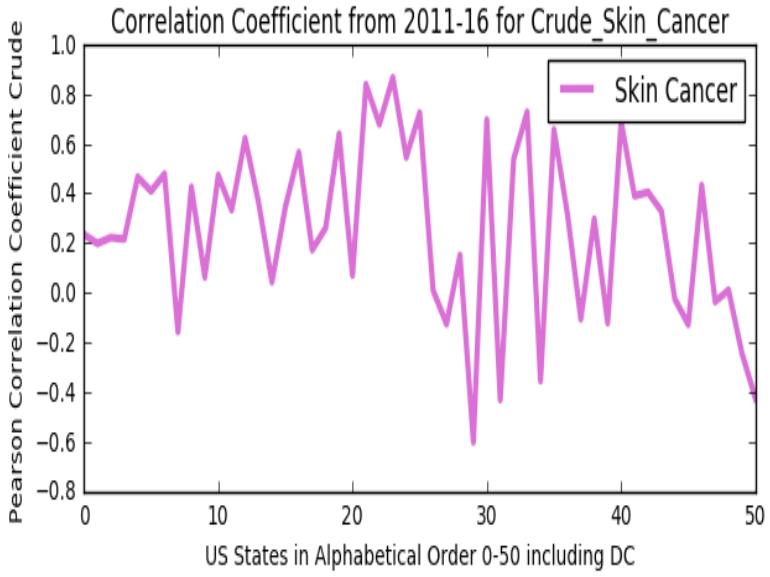

Figure 6 explains the Correlation between our Ground Truth (Crude) and Google trend data for search term 'Skin Cancer'. This correlation is presented here as State wise (all US States and Washington DC) and is aggregated temporally (2011-2016). Please also note that US states are sorted alphabetically and their corresponding rank in this sorted order is presented here instead of the name.

Figure 7

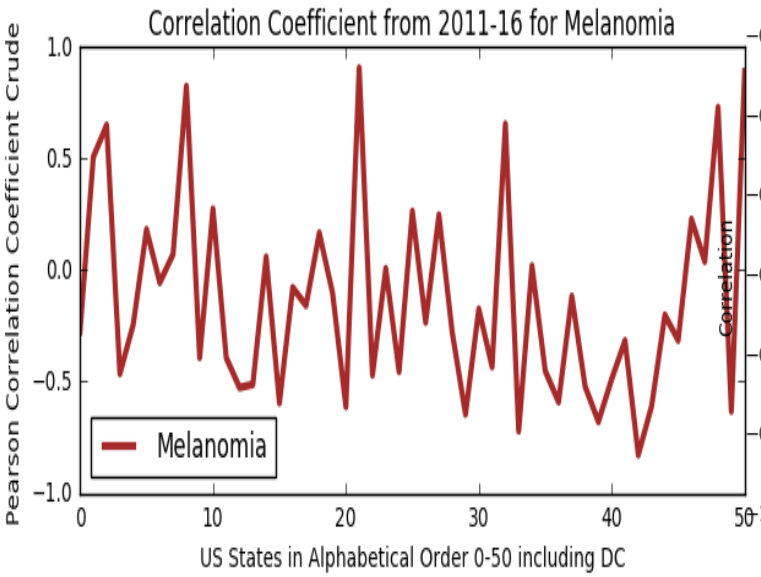

Figure 7 explains the Correlation between our Ground Truth (Crude) and Google trend data for search term 'Melanoma'. This correlation is presented here as State wise (all US States and Washington DC) and is aggregated temporally (2011-2016). Please also note that US states are sorted alphabetically and their corresponding rank in this sorted order is presented here instead of the name.

Figure 8

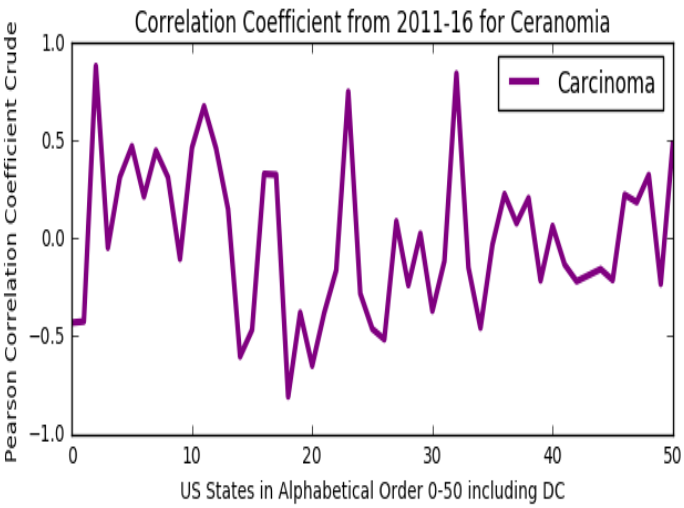

Figure 8 explains the Correlation between our Ground Truth (Crude) and Google trend data for search term 'Carcinoma'. This correlation is presented here as State wise (all US States and Washington DC) and is aggregated temporally (2011-2016). Please also note that US states are sorted alphabetically and their corresponding rank in this sorted order is presented here instead of the name.

Figure 9

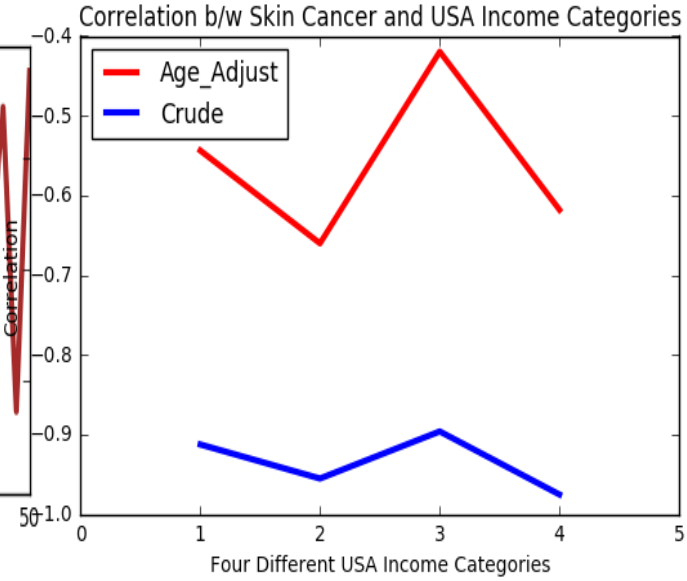

Figure 9 explains the Correlation between our Ground Truth data and four distinct US Income categories. These Income categories respectively are 'Male Full Time', 'Female Full Time', 'Male Seasonal Worker' and 'Female Seasonal Worker'.

Figure 10

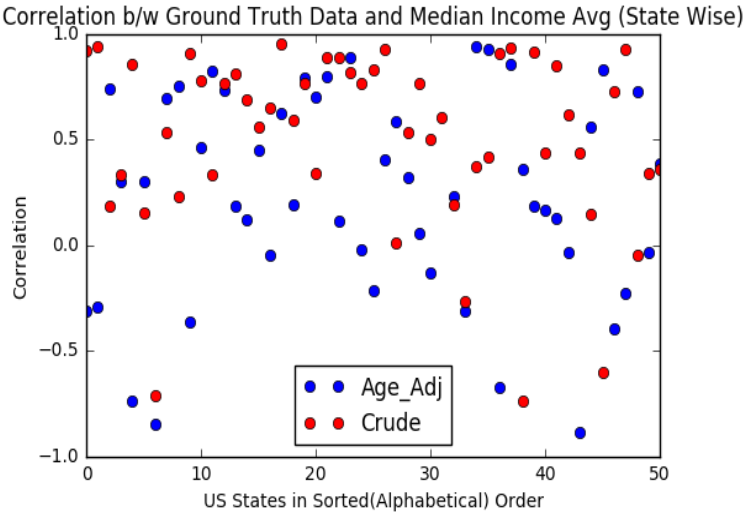

Figure 10 explains the Correlation between Ground Truth Data and Median Annual US Income. This correlation is represented State wise (States are sorted in an alphabetic fashion and their rank is mentioned here).

Figure 11

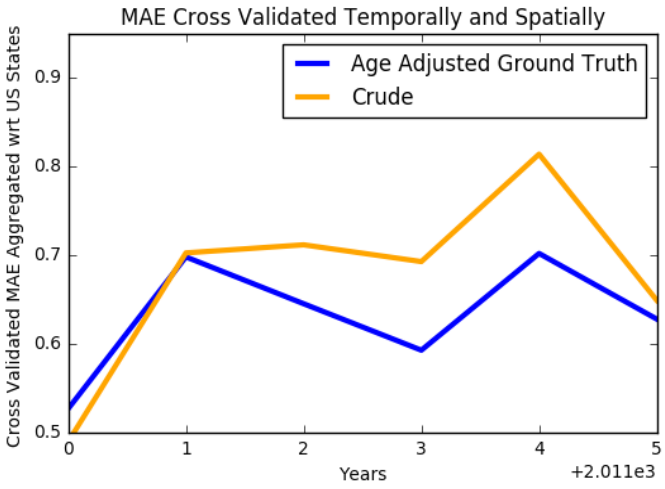

Figure 11 explains the prediction performance of Lasso Regression. The input features are extracted from Google Trend Data for three search terms namely 'Skin Cancer', 'Melanoma' and 'Carcinoma'. Th output variables are Ground Truth Skin Cancer data. The model selection is based on temporal and spatial cross validation. The performance is measured on the criterion 'Mean Absolute Error' aka 'MAE'. The 'MAE' is calculated six times corresponding to each different year (2011-16).

Figure 12

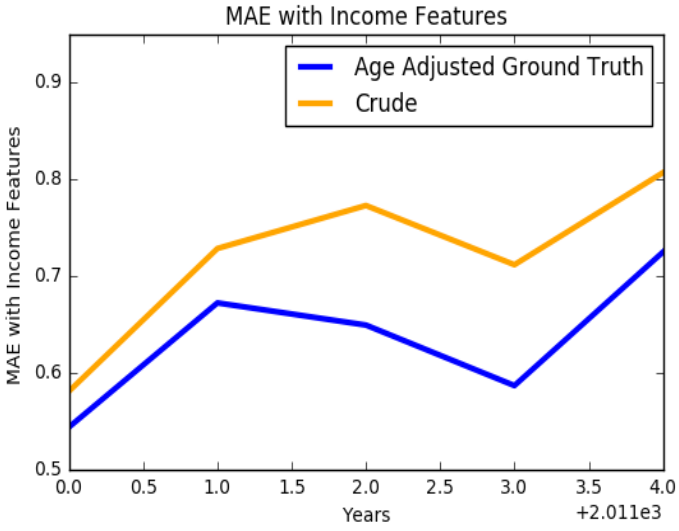

Figure 12 explains the prediction performance of Lasso Regression. The input features are extracted from Google Trend Data for three search terms namely 'Skin Cancer', 'Melanoma' and 'Carcinoma' and Median Annual Income in US. Th output variables are Ground Truth Skin Cancer data. The model selection is based on temporal and spatial cross validation. The performance is measured on the criterion 'Mean Absolute Error' aka 'MAE'. The 'MAE' is calculated five times corresponding to each different year (2011-15).

Figure 13

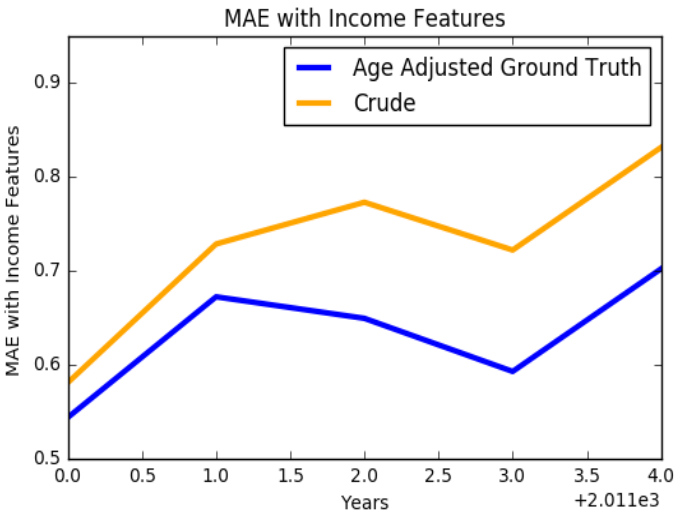

Figure 13 explains the prediction performance of Lasso Regression. The input features are extracted from Google Trend Data for three search terms namely 'Skin Cancer', 'Melanoma' and 'Carcinoma', Median Annual Income in US and the Percentage of Uninsured people in the US. Th output variables are Ground Truth Skin Cancer data. The model selection is based on temporal and spatial cross validation. The performance is measured on the criterion 'Mean Absolute Error' aka 'MAE'. The 'MAE' is calculated five times corresponding to each different year (2011-15).

VI. Discussion and Conclusion

.

We try to conclude our understanding achieved by working on this project. After Careful review of Correlation Coefficient and MAE, we do not believe that ideal Machine Learning work can be done to understand the specific disease under investigation 'Skin Cancer' with the help of the features mentioned in the document. The Correlation coefficient is very low. In fact, we believe it's too low to be considered significant. The 'MAE' although shows that we can learn from this data, it also shows that there's a lot of under fitting. More specifically, we tried different Machine Learning algorithms including Support Vector Machines, Random

Forest, Ridge Regression and Ordinary Least Square and even the training error is significant. This means that there's not a lot to learn from this data or in other words, if we choose the most obvious Machine Learning algorithms from literature, it's likely that they'll perform very bad. Thus, we conclude that we do not believe that a Specific pattern exists inside this data set (Google Trend Data) which can be learned easily. Of Course, if we use different data (different keywords, different years), the learning may improve at times.

REFERENCES

[1] US Center for Disease Control and Prevention. Available at https://www.cdc.gov/brfss/

[2] Google Trend Data for US. Available at https://trends.google.com/trends/

[3] Median Annual Household Income Statewise. Available at https://en.wikipedia.org/wiki/List_of_U.S._states_by_income

[4] Percentage Uninsured person in USA statewise. Available at https://en.wikipedia.org/wiki/Health_insurance_coverage_in_the_United_States

[5] Sickit Documentation Available at http://scikit-learn.org/stable/