INFOSYS 722 DATA MINING AND BIG DATA
Assessment 4, Flu forecasting

Quentin MAS (tmas060)

September 2019

1 Step 1 : Business and Situation understanding

1.1 Objectives of the situation

The objective of this situation is to predict the spreading of Flu in the U.S.. the flu as an important impact on the economy the estimation of the annual economic costs of influenza varied from $13.9 thousand to $957.5 million across US counties(1). And a pandemic Influenza could cost to the united State $71.3 billion to $166.5 billion.(2)

Being able to forecast Flu would help to manage stock of drug (knowing which state is going to need more of them), but also to organize vaccination and prevention campaigns. All of this will reduce the cost of influenza for the economy and help to prevent a pandemic Influenza.

1.2 Assessment of the situation

Influenza, is an infectious disease caused by an influenza virus. It is usually spread through the air from coughs or sneezes.
On average, about 8% of the U.S. population gets sick from flu each season.
Most people will recover completely in about one to two weeks, but some people are at high risk of developing serious flu-related complications if they get sick (people 65 years and older, pregnant women, and children younger than 5 years, people of any age with certain chronic medical conditions such as asthma, diabetes, or heart disease). Because Influenza is much more than a common cold, and can be dangerous for some people there is a need to prevent it spreading.

There is a lot of data on the evolution of influenza in the United States available on the CDC (Centers for Disease Control and Prevention) website. Other data on the United States (populations, states...) are available on the U.S. Census Bureau website.

The main risk is that there is some unexpected error in the data (poor quality, wrong estimation of the geographical spread) that will create an error in the data set.

In the case of data of poor quality there is not much I can do. But this estimation are made by a trustworthy organization. What I mean is there may be some error, but not many. Because of this assumption I will not delete data because I assume it is of poor quality.

If the initial results are less dramatic than expected I will try to find other external data to add in the model.

1.3 data mining objective

The goals are to use the data from the previous flu-seasons to forecast influenza activity in each state of the U.S. for the next week.

1.4 Project plan

The project plan is firstly to understand the data and to find other relevant data for each state. Then merge all this data and clean them. With this new dataset an effective model can be researched. Finally, we can evaluate the performance of the chosen model.

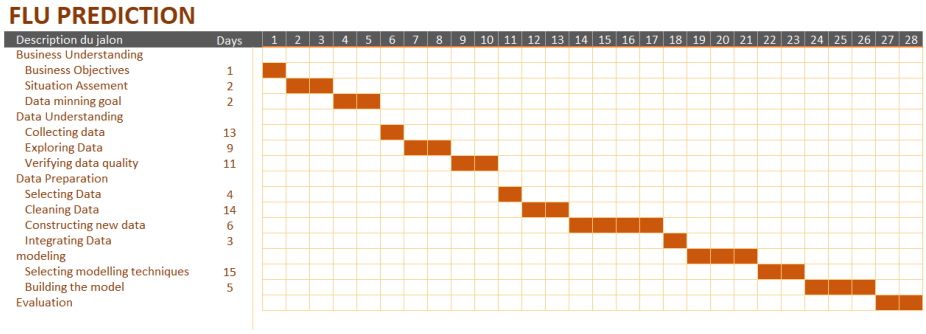


Figure 1: Distribution of the activity for each week

2 Step 2 : Data Understanding

2.1 Initial data

Influenza activity : Each week, for each state, the CDC (Centers for Disease Control and Prevention) makes an estimation of the geographic spread on influenza(3). The spreading in each state is classified into five classes (No Activity, Sporadic, Local Activity, Regional, Widespread). This dataset contain not only the 50 American states, but also territory, such as Guam which are haven't a big population and which may need to be removed as part of the data preparation process. This estimation is made only during the flu season (from the week 40 of one year to the week 20 of the next one)

Population : The Census Bureau produces each year an estimation of the population for each state. (4)This estimation is part of the Census Bureau's Population Estimates Program (PEP). To do this estimation the census bureau take the population from the previous year and add to this number the number of birth and the number of people who moved in the state, subtract the number of death and the number of people who moved out of the state. their is a few issue with this method, first if there is an error during one year the same error will still be there during the next year estimation. Also this method doesn't take into account the illegal immigration which can represent up to 7% of the

population(9).

Household income : mean household income for each state. This is an estimation made by the Census Bureau from several major national household surveys and programs(5). This data come from a survey so they are subject to who have been interrogated and how people have answer (some may have increase a little their income).

Assurance : this dataset gives the number of people having or not a health assurance. Before 2013 this data where obtained from the Current Population Survey and since 2013 they are obtained from the American Community Survey. The fact that this data come from two different surveys may be a problem. But the two survey are made by the Census Bureau so it should be good. (6)

Area : Area of each state in km^2. This dataset as been made in 2010 and come from the Census Bureau. Area

Typevirus : This data set saw what type of flu (type and lineage) is spreading in the US. The estimation is made by the CDC each week for the whole country (not for each state)(8). This data are extracted from the Influenza Hospitalization Surveillance Network. In this network hospital and laboratory report in detail every case of influenza. An issue here can be the people which don't go to an hospital or diagnosis error.

2.2 Description of the data

The Flu activity dataset contains more than 29 000 data points between 2003 and 2019. It contains five fields : STATENAME, ACTIVITYESTIMATE, SEASON, WEEKEND (date of the weekend).

The population dataset contains data of the estimated population of each state between 2010 and 2018.

The Households income dataset gives annual information between 1984 and 2017.It contains three fields : STATENAME, INCOME, YEAR.

The assurance dataset gives annual information between 2002 and 2017. It contains five fields : STATENAME, HealthCare_All (percentage of the population which have an Health Insurance), HealthCare_Under18 (percentage of the population under 18 years old which have an Health Insurance), HealthCare_Under65 (percentage of the population under 65 years old which have an Health Insurance), HealthCare_Over65 (percentage of the population over 65 years old which have an Health Insurance).

The area contain the area of each state in km^2. The data set I used i only a part of the original one you can find on the CDC web site which contain much more geographical information. It contains two fields STATENAME and AREA

The virus type dataset give information about what kind of virus is spreading in the US. It contains eleven fields : WEEK, STATENAME, TypeA_subH3, TypeA_subH1N1, TypeA, TypeA_subU, TypeA_subH1, A_subH3N2, TypeB, TypeB_subVL, TypeB_subYL.

As you can see the majority of the variables have a name similar to Type*_sub*, they give the number of case recorded for each type of flu. The first part of the name (Type*) give the type of virus (A or B) and the second part (sub*) the sub-lineage.

Some data sets have the field YEAR meanwhile the Flu dataset has the parameter SEASON. To make the merge easier I chose to Change the Year in the associated season (for instance '2018' became '2017-18').

Some values are categorical ('SEASON', 'STATENAME' 'ACTIVITYESTIMATE', 'WEEKEND') all the others are numerical.

2.3 Explore the data

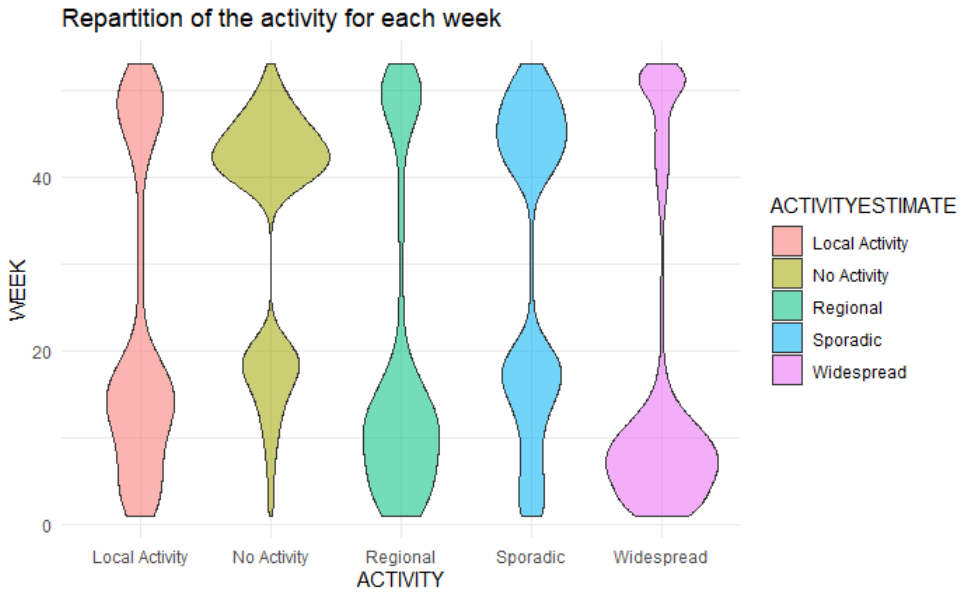


Figure 2: Distribution of the activity for each week

This graph represents during which week an activity is more present. For instance the Flu is usually widespread at the beginning of the year, during the same weeks the States where there is no activity are really scarce. And the closer we get to summer, the more states have a local activity, a sporadic activity or no activity. This simply shows that depending on the week, states have a high probability to have high or low activity.

We can also see that there are some data during summer (even if the CDC doesn't write any reports between the week 20 and 40). This will be explained in the next part.

This graph show the distribution of the flu activity in each state for all the dataset. We can see it isn't uniform, some state are more often in a certain activity than other. for exemple in the state of New York the flu activity is Widespread in nearly half of the dataset (260 times) . Meanwhile the activity in the District of Columbia was only three

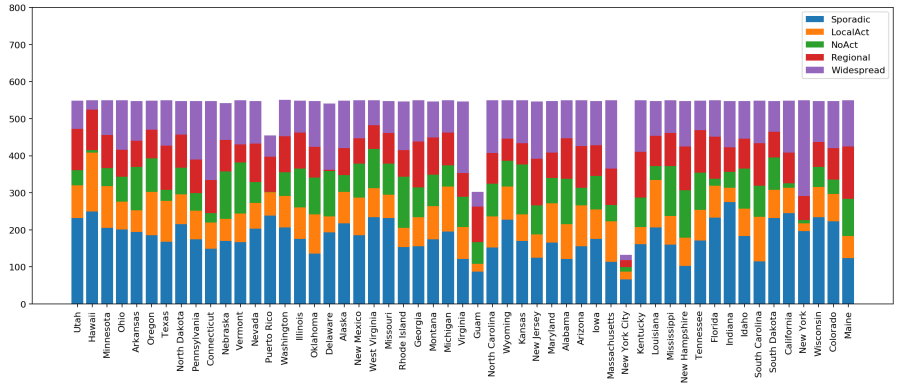


Figure 3: Distribution of the flu activity for each State

times widespread.

2.4 data quality

There are few problems with the Flu dataset.

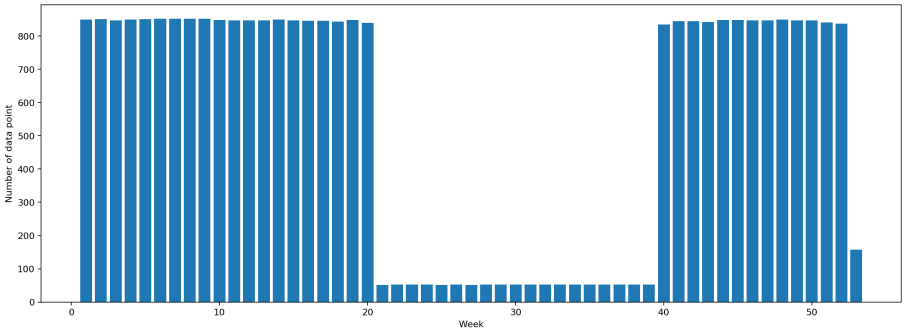


Figure 4: Distribution of the data for each Week

As you can see, there is an important period in summer during which there are almost no data. This may not be problematic because the flu isn't very active during this period. In fact the CDC doesn't make estimation during this time of the year for this reason. Still there is some data available. This data come from the years 2008-2010, during this year the HN flu was really active that's why they continued to make estimation during this two summer. There is also more data during the last week this is just because some years have 53 weeks.

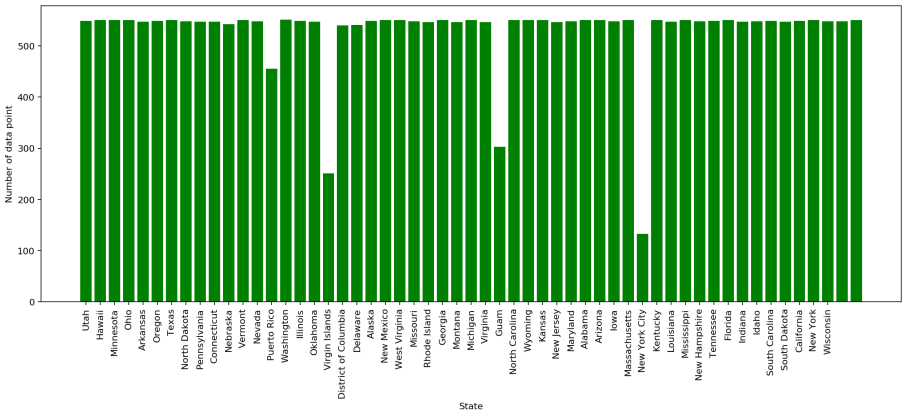


Figure 5: Distribution of the data for each State

Another problem is the number of data points by state. As you can see, there are four states (Georgia, New York, Pennsylvania and Vermont) where there are not as many data point as there are in the other states. This could be a problem for the data mining step and to implement some model.

The others datasets don't have missing data or data errors. But the household income dataset and the assurance dataset end in 2017, the population dataset end in 2018. And we want to do prediction in 2019. This problem can be partially solved by first making a prediction of this variable.

3 data preparation

3.1 select the data

Some data of the Flu activity dataset such as the weekend are useless. The weekend attributes give the date when the estimation was made, but we already have the week number which is more interesting.

During most summers there are no data this field could be completed by "No Activity" because the reason behind these missing data is that is not really active during summer. But I choose to delete this row because after looking to the last week of each season and the first one I saw that most states were in the state Sporadic and some in local activity(as you can see on the following graphs which represent the activity of flu in the US during the week 40 and the week 20).

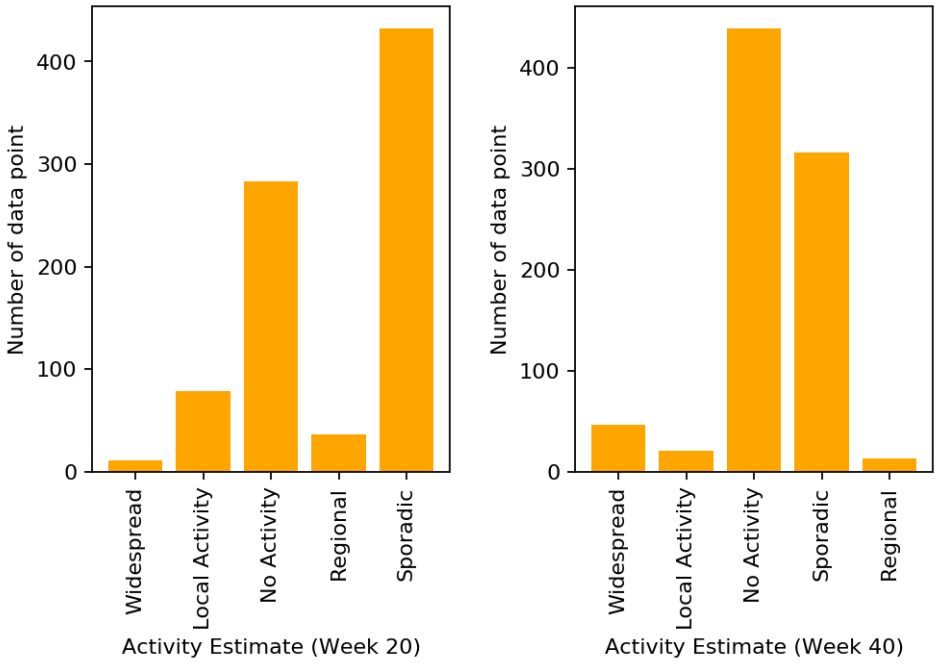


Figure 6: Activity for the week 20 and 40

Some territories such as Guam have less data than others and have less population, which make them less interesting. I decided to delete these territories from the dataset. I also choose to delete all the state/ island outside the US mainland (Alaska, Hawaii). These territories have less connection and so I chose to work with only the Contiguous United States.

The population dataset only goes from 2010 to 2018 so I choose to select only the data point after the season 2009-10

3.2 clean the data

The data in the different data set are supposed of good quality as they come from national institute. But there is some problem in the Virus Type data set. There is missing data for the two lineages of type B (TypeBsubVL, TypeBsubYL). What I think is before 2014 they were only counting type B virus without making the difference between the lineage. If you look at the following graph you can see the number of flu type B (in red) decrease in 2014. But the global number of cases of type B continued to grow.

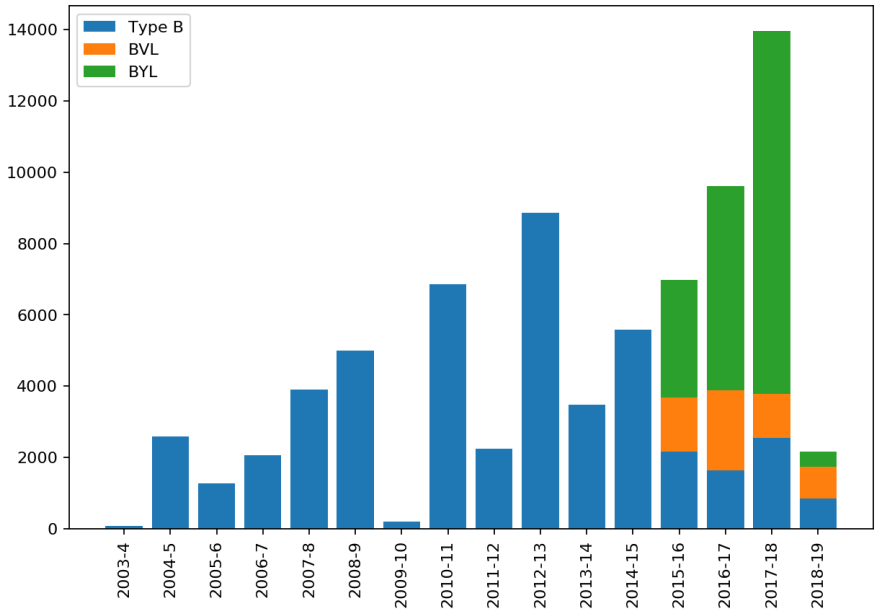


Figure 7: Influenza type B

Because of that I chose to regroup this three feature into one (TypeB) by considering the missing value equal to zero.

3.3 construct the data

To predict the influenza activity you need to know what was the activity of Influenza during the previous week. So temporal attribute must be created by deriving others attributes.

Here is the feature I choose to add :
- activity of a state during the previous week. The goal of this feature is to give a basis for the estimation. A disease doesn't infect a whole state in an instant it evolve.
- the number of states that was in each activity class during the previous week. This new feature are named NRegional, NLocalAct, NSporadic, NWidespread. The idea behind these features is to represent the global state of the spreading of flu in the whole country. If the flu is widespread in a lot of state it may mean the flu is highly spreading.

3.4 Integrate various data source

First, there is not only one kind of Flu, each one is different and spread at a different speed. So, data about the proportion of each of this Flu must be added. That's why I integrate the Type virus data set
I also choose to integrate data about each state. More precisely about the population (number of inhabitants, density, part of the population being covered by an insurance, income).

3.5 Format the data as required

After Integrating different data source I saw that some data weren't in the correct format. All the data from Health were seen as Object by python so I changed the type of the variables HealthCareAll, HealthCareUnder18, HealthCareUnder65, HealthCareOver65 to float.
There are also some columns are strings (for instance STATENAME or SEASON), so I converted them into numeric data using the function OneHotEncoder().

4 Data transformation

4.1 reduce the data

Let's first look at the correlations to see if any variable is to much correlated an other one.

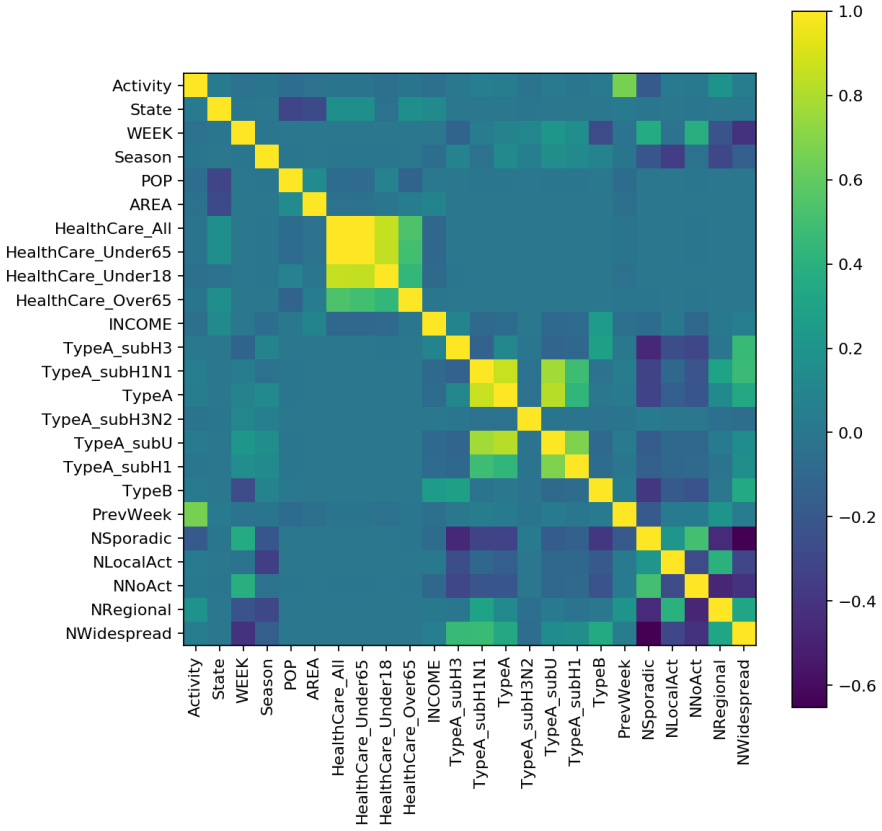


Figure 8: Correlation between the variables of the dataset

You can see there isn't a value that is highly correlated with the target (*Activity*). The only column with a high correlation factor is PrevWeek with 0.7. In this case I don't think having a correlation factor of 0.7 is a problem because this feature is really important as it describes the activity during the previous week, which is an essential

temporal information.
There are also some variables that have a correlation factor close to zero. In a first time I won't delete them, but I may choose to do so after a few iterations.

If you look at the correlation between other variables, you can see some variables are highly correlated. It is especially true with the Healthcare data. The variables *HealthCare_under65* and *HealthCare_Under18* are highly correlated with *HealthCare_All* (correlation ¿ 0.8). A third similar variable (*HealthCare_Over65*) has a strong correlation with *HealthCare_All*, but it's reasonable and it's about a part of a population highly sensible to Flu. Seeing this I chose to delete the variable *HealthCare_under65* and *HealthCare_Under18* and to keep *HealthCare_All* and *HealthCare_Over65*.
Other variables are highly correlated such as *NSporadic* and *TypeA_subH3*, but since I don't see any reason for this correlation I chose to keep both variables.

After a first iteration I obtained the following graph for the variable importance:

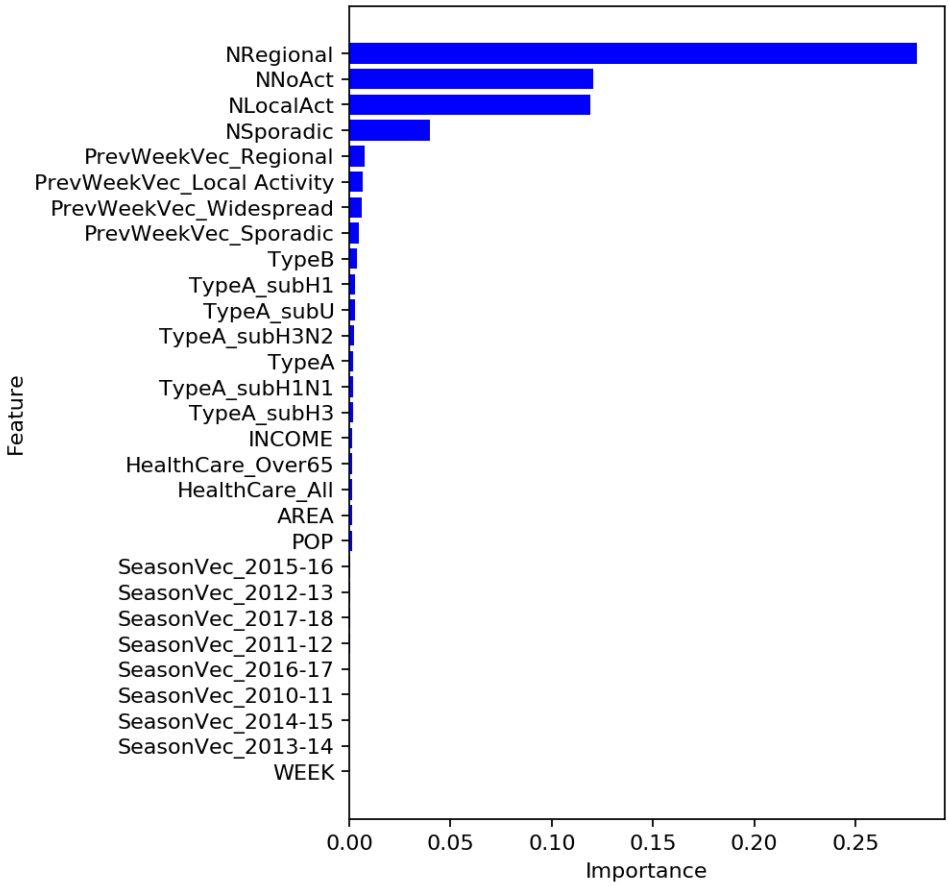


Figure 9: Iteration 1 : Feature Importance

As you can see the importance of some variables is null. So I decided to delete the following variables : *SEASON*, *STATENAME*, *TypeB*. *WEEK* also as an importance near

to zero yet I chose to keep it because it give an important temporal information, the flu is a seasonal sickness (the flu is more active during certain week than during others (Figure 2)

So, after a second iteration I obtained the following graph for the feature importance :

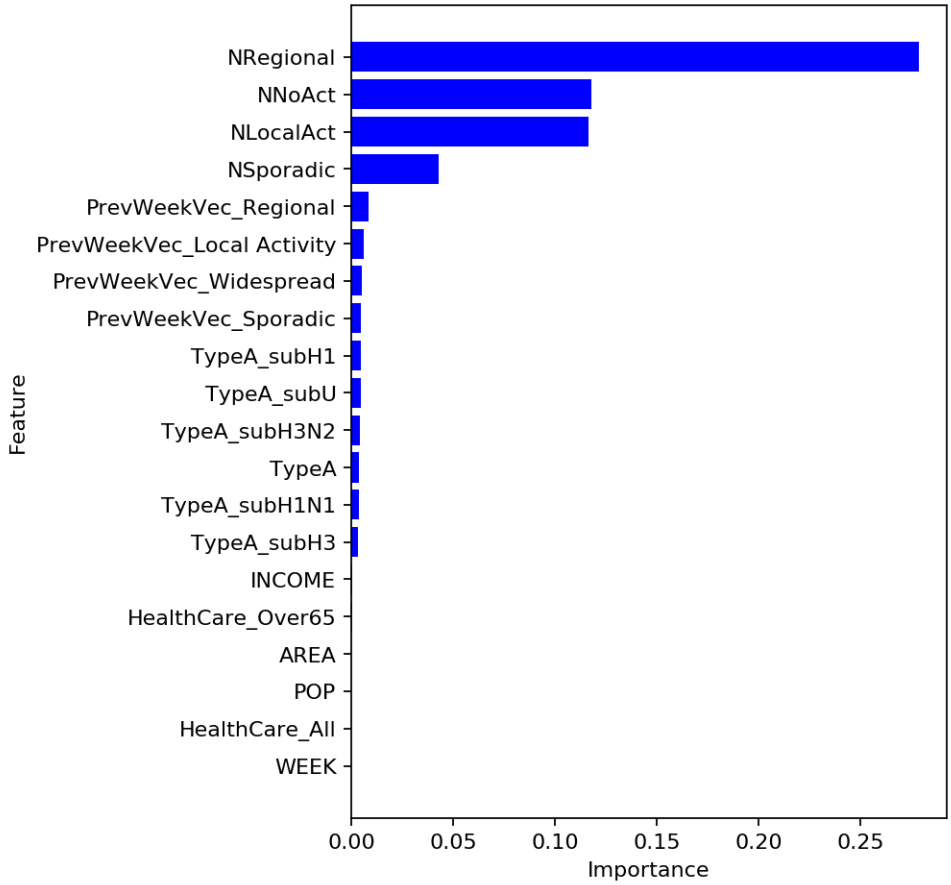


Figure 10: Iteration 2 : Feature Importance

Because all variables seems or could be useful I chose to keep them like this.

4.2 project the data

The distribution of NWidespread is given in the following graph.

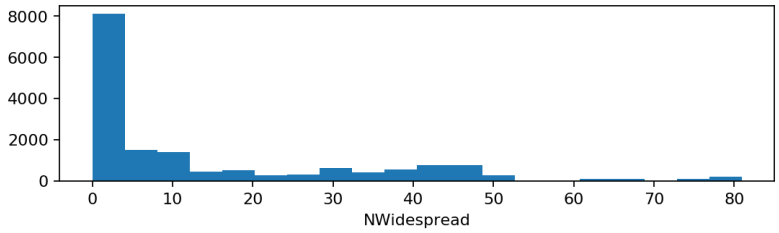


Figure 11: Distribution of NWidespread

As you can see this feature usually have the value 0. To give more importance to the other value of NWidespread I choose to apply a log transformation. The resulting distribution is the following graph.

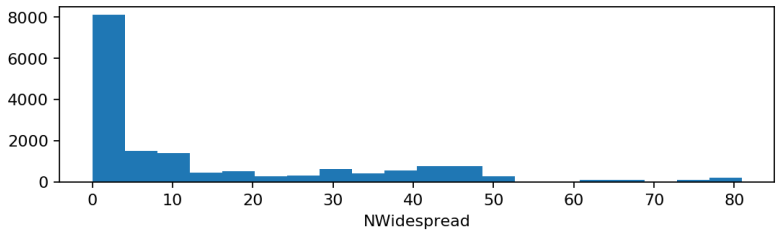


Figure 12: Distribution of NWidespread after log transformation

5 Data mining method selection

I am looking to predict what will be the flu activity during the next week. The data I have, gives me information about the activity of flu. Because I have this labelled data I will be looking into Supervised learning Algorithm.
As I said I want to predict the influenza activity, this can be done in a variety of ways. I could try to predict the number of people getting sick each week. I could also try to classify the spreading of the flu in different class indicating the severity of the spreading.

I have information about the number of people getting sick each week with the type virus dataset. But I think predicting the number of sick people can be hard. Moreover I don't have the number of sick patients in each state only in the whole country. And as I said about in 2.3 the spreading of the flu can change a lot depending of the state.

On the other hand I have the Flu activity dataset which gives me the flu activity in each state. This data contain six different levels and could be used with a classification method.

At first glance it seems this is a prediction problem, but whit this dataset I think a classification methods will perform better.

6 Data mining algorithm selection

6.1 analysis of DM algorithms

Different algorithms can be used for this problem. Neural networks can be a good model in this case, they are good to find and use complex relations between the different variables. Moreover, they have been proven to work very well in a lot of different problems. However, they are well know for being really hard to understand and interpret.

Decision Tree can also perform well on this kind of problem. The results of this method are more understandable than neural networks.

Naive Bayes classifiers which are a family of simple probabilistic, multiclass classifiers based on applying Bayes' theorem. This method is also well understandable.

6.2 model selection

What the user is interested in, is a good prediction, but the model need to be understandable in order to fight more easily the spreading of the flu.

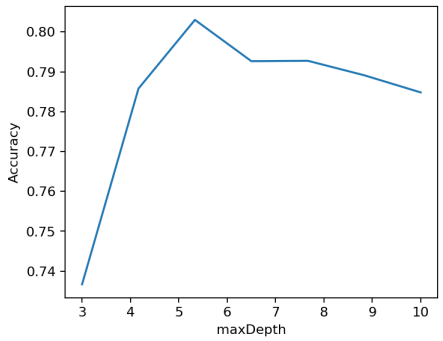
Because the model need to be understandable, I choose to not use neural network even if they could have performed well on this problem. I choose to use Decision Tree and Naive Bayes classifiers which I think is a good tradeoff between performance and comprehension. But Naive Bayes classifiers are based on applying Bayes theorem with strong independence assumptions between every pair of features, which is not the case here with every variable (4.1). Anyway, keeping that in mind I chose to still continue with this algorithm.

6.3 build the model

After trying both algorithm I saw the Naive Bayes classifiers didn't perform well. I manage to obtain a accuracy of 0.78 with one neighbor. Meanwhile,I obtained an accuracy of nearly 0.8 with the decision tree. Their isn't a big difference but since all features aren't independent I chose to continue only with decision tree.

A decision tree has a few parameters that can be tuned in order to improve the prediction of the model :
maxDepth This parameter give the maximum depth of the tree. the Higher this parameter is the the more complex the tree will be. If this number is to big it can result in overfitting.
maxBins for faster tree calculations the algorithm split the data in bins to make calculation on them and not on all the dataset. This parameter give the maximum number of bin that can be created. With a high number of bins the algorithm will be faster and with a slow number of bins it may be better but will require more computing power. Since I'm limited in computing power I have to find a sweet spot for this parameter.
minInstancesPerNode : the minimum number of samples to split a node. With a low number of sample the tree may grow deeper.

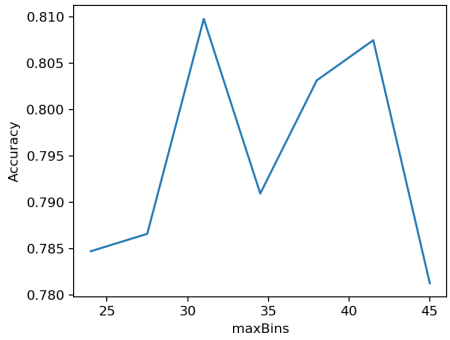
Let's try different values for maxDepth:



This graph show the accuracy in function of maxDepth. As you can see the accuaracy is rising up to 0.8 and then decrease. Since the maximum accuracy is obtained with a maximum depth of 5 I chose to use maxDepth = 5.

Figure 13: Tuning maxDepth

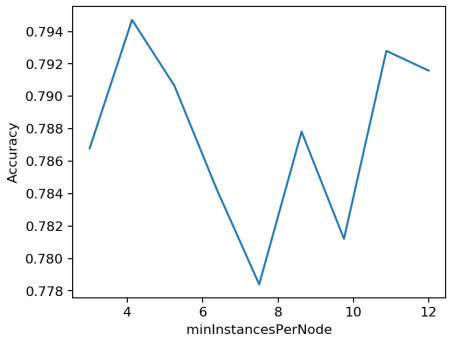
Let's do the same with maxBins :



As you can see the accuracy is maximised around 30 and 40. Since the computing power of the EC2 instance is very limited I chose tu use maxBins = 40.

Figure 14: Tuning maxBins

Finally let's tune minInstancesPerNode :



I could have chose minInstancesPerNode equal to 4 or 11, but with this parameter equal to 11 the tree would be smaller which mean less computing power and a tree easier to understand. Because of this reason I chose minInstancesPerNode = 11

Figure 15: Tuning minInstancesPerNode

7 Data Mining

7.1 test design

In order to evaluate the performance of different model I decided to split the dataset in two parts. A train set (90% of the database) to train the model and a test set (10%) to evaluate the model. The split could be done either randomly or using the date (select the last week of the data set for the test set). I choose to do the split randomly because if i used the second possibility I may select a model that predict well the kind of flu we have currently but will be bad at predicting a pandemic flu as H1N1.

7.2 patterns

The different Activity Classes have an order, so let's see if when the algorithm is wrong if it was because it had chosen the next classes (for instance let's check if it didn't said widespread when they were no activity).

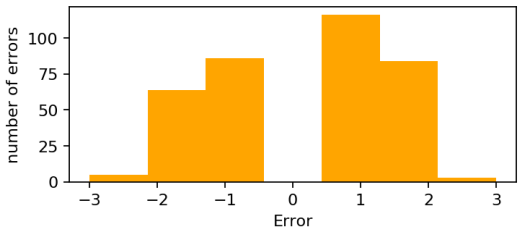


Figure 16: error between the classes

This graph represent the distance between the predicted classes and the true class (if the predicted activity is "No Activity" meanwhile the true Activity is "Local Activity" then the distance will be two). The test set has a size of 1600. This graph doesn't present the good prediction.

We can see the number of errors is decreasing with the size of the error, which is good. but on the other hand the model is often completely wrong ($|error| > 1$). Finally we can see the graph is not symmetric, there is more error on the right side of the graph than on the left side.

If we look at the importance of variable (Figure 17) we can see that the main variables are *PrevWeek* and the features that describe the flu activity in the whole country (*NRegional*, *NNoAct*, *NLocalAct*, *NSporadic*). Surprisingly the week isn't that munch important.

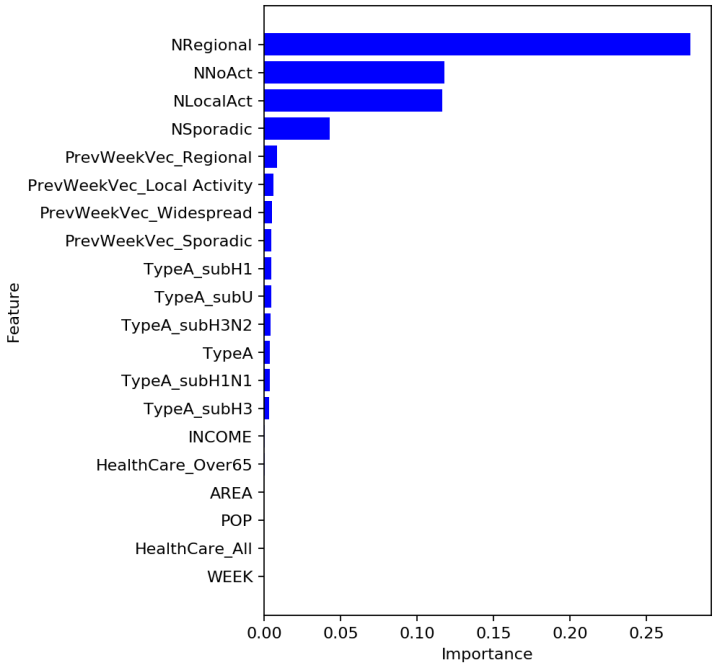


Figure 17: Feature Importance

8 Evaluation

8.1 patterns discussion

If the algorithm is completely wrong about the flu activity ($|error| > 1$) it is because sometimes the flu spread quickly and a state can go from sporadic to regional in a week. Maybe more temporal feature could help to solve this problem.
Also the lack of symmetry show the model more often underestimating the flu activity (The graph has been obtained by computing the true activity minus the predicted one).

PrevWeek is the most important variable because the algorithm has to know what was the activity before to predict it. The features that describe the flu activity in the whole country (*NRegional*, *NNoAct*, *NLocalAct*, *NSporadic*) are also really important because they are giving a general information about the spreading of flu. I think it is beacause of this variables that week has a low importance. In fact variables like *NRegional* give a similar information to week but with more precision.

8.2 model performance

At the end the model give of 0.82 which is not as good as expected. Yet the accuracy may be improved by creating new temporal feature about the spread of Flu

8.3 Multiple iteration

To ensure the quality of the model I tried to train the model on ten different train set and to test them on ten different test set. The result are given in the following graph.

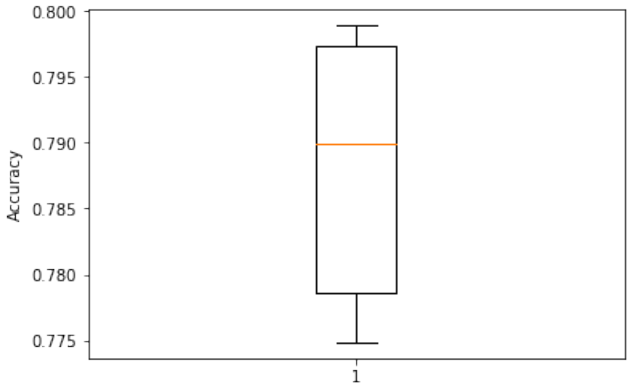


Figure 18: Accuracy of the model on multiple iteration

This boxplot show the the true accuracy of the model is more of 0.79 and is include between (0.77 and 0.8). At the end the mean accuracy of the model isn't very good but the deviation is quite good.

References

[1] Liang Mao, Yang Yang, Youliang Qiu Yan Yang (2012).Annual economic impacts of seasonal influenza on US counties: Spatial heterogeneity and patterns. https://ij-healthgeographics.biomedcentral.com/articles/10.1186/1476-072X-11-16.

[2] Meltzer, M. I., Cox, N. J., Fukuda, K. (1999). The Economic Impact of Pandemic Influenza in the United States: Priorities for Intervention. Emerging Infectious Diseases, 5(5), 659-671. https://dx.doi.org/10.3201/eid0505.990507

[3] Influenza Activity, https://gis.cdc.gov/grasp/fluview/FluView8.html

[4] Population by State, Census https://www2.census.gov/programs-surveys/popest/datasets/2010-2018/state/asrh/

[5] Household Income, Census https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-household

[6] health Insurance, Census American Community Survey https://www.census.gov/data/tables/time-series/demo/income-poverty/cps-hi/hi-05.html https://www.census.gov/data/tables/time-series/demo/health-insurance/acs-hi.html

[7] State Area, Census https://www.census.gov/geographies/reference-files/2010/geo/state-area.html

[8] Virus Type, CDC https://gis.cdc.gov/grasp/fluview/flu$_b y_a ge_v irus.html$

[9] U.S. unauthorized immigrant population estimates by state, Pew Research Center, 2016 https://www.pewresearch.org/hispanic/interactives/u-s-unauthorized-immigrants-by-state/

I acknowledge that the submitted work is my own original work in accordance with the University of Auckland guidelines and policies on academic integrity and copyright. (See: https://www.auckland.ac.nz/en/students/forms-policies-and-guidelines/student-policies-and-guidelines/academic-integrity-copyright.htmlLinks to an external site.).

I also acknowledge that I have appropriate permission to use the data that I have utilised in this project. (For example, if the data belongs to an organisation and the data has not been published in the public domain then the data must be approved by the rights holder.) This includes permission to upload the data file to Canvas. The University of Auckland bears no responsibility for the student's misuse of data.