Respiratory viral DREAM challenge 11: Ensemble of predictions Project Report

Introduction :

Respiratory Viruses affect day to day life of people and may also lead to serious secondary diseases, even cancers sometimes. Some people who are exposed to virus are completely immune to those viruses. In that case, by obtaining the gene expressions before and after virus exposure/attack on individuals, it will be easy for us to protect the non-immune individuals from further attacks. We were presented with codes submitted from various teams across the world for finding the predictors related to symptomatic changes in an individual. The goal is to provide an Ensemble model based on the submitted models that can be used for future predictions.

Workflow Diagram:

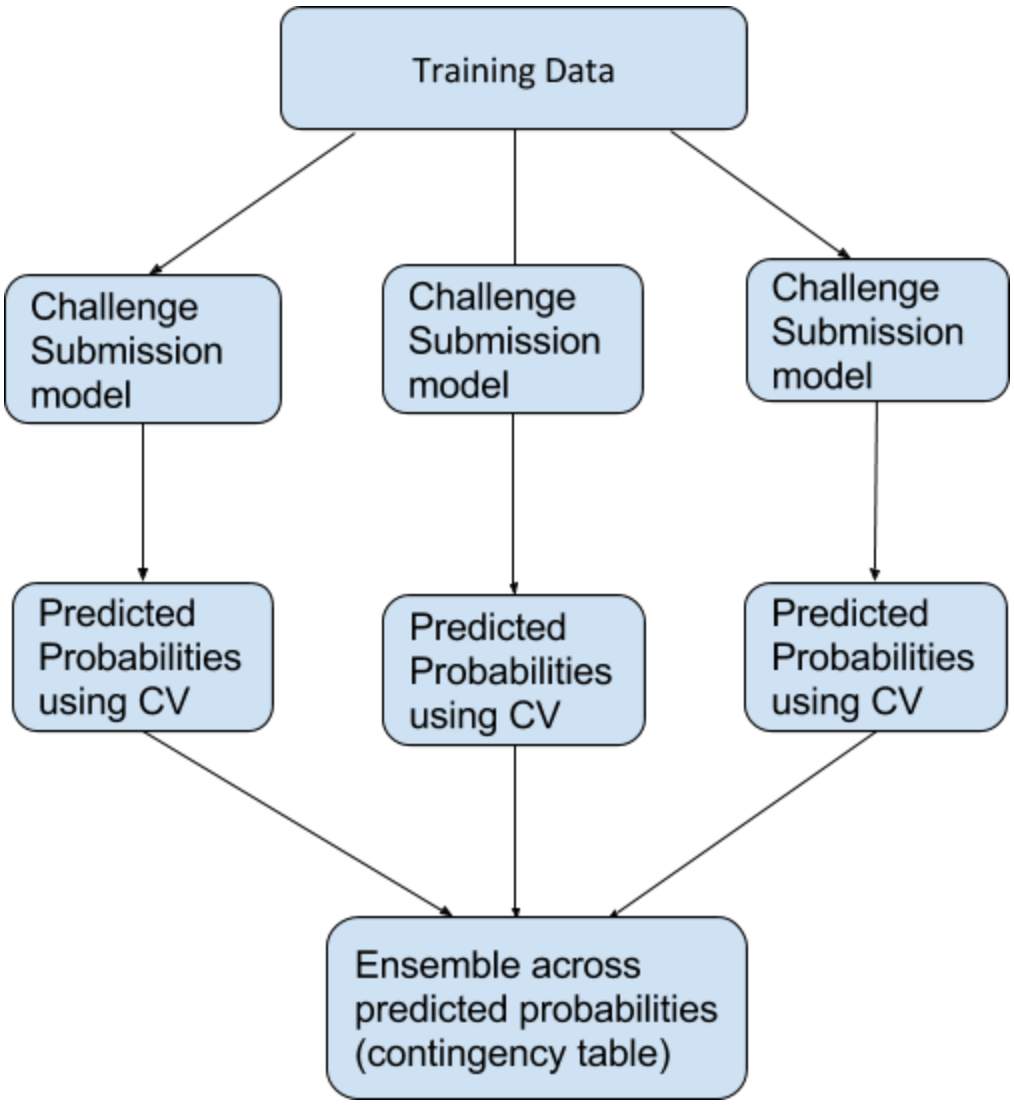

Fig 1.1 Workflow diagram

Data overview :

4 Respiratory viruses are taken into consideration (H1N1, H3N2, RSV, Rhinovirus) and 7 experiments. Healthy volunteers are exposed to 4 viruses and gene expressions are studied before and after the virus influences. This leads us to two sets of training data - Viral_Challenge_CLINICAL data and Viral_Challenge_Expression data. These are the clinical and normalized expression data. The Expression data has 22277 rows and 2371 columns. The rows of the normalized Expression data denote the observations(genes) and the columns of the Expression data denote the variables (experiments). Similarly in the Clinical data set, there are

2371 rows and 12 columns. The rows represent the experiments and the columns are the 12 variables whose explanation is as follows:

LOGSYMPTSCORESC3	**log10(max symptom score +1)**
TIMEHOURS	**Time of gene expression profile (hours)**
SAMPLEID	**Gene expression sample ID**
CEL	**CEL file name**

Data Exploration & Cleaning:

Variable Name	**Variable Description**
STUDY_ID	**Study name**
SUBJECTID	**Unique patient ID**
AGE	**Patient age**
GENDER	**Patient gender (Male or Female)**
EARLYTX	**= 1 if patient received early oseltamivir,** **= 0 if patient received oseltamivir at day 5,** **or NA for cohorts other than DEE5 H3N2**
SHAM	**= 1 if patient received sham exposure,** **= 0 if patient received viral exposure,** **or NA for cohorts other than Rhinovirus Duke**
SHEDDING_SC1	**= 1 if patient exhibited viral shedding,** **= 0 if viral shedding not observed**

SYMPTOMATIC_SC2	**= 1 if max symptom score >=6,** **= 0 if max symptom score < 6**

Table 1.1 Data Exploration

To perform the Ensemble methods and Feature Selection on the dataset, the data needs to be cleaned keeping only the relevant information and removing the noise elements. No missing data imputation method is satisfactory if we have too many missing data.
The dataset that we are using is already cleaned from the previous submissions. We have hence not performed this step from our end.

Feature selection:

We used t.test as feature selection from previous submissions. We set the p.value threshold to a very low value 0.00005. We used only those data values which satisfy the condition, i.e., p.value lesser than threshold and use them as train and test data after doing K fold Cross validation. By doing this we selected the best features. Also we used t.test to avoid overfitting. This part of the code was presented to us to find the models and to find the Ensemble model. This is not done by us.

Classification:

We trained our data using models KNN, Lasso and SVM from 3 submissions ViResPred, Agnita and Shotsy respectively .
We have used the below submissions for building models for producing the final Ensemble model:

Model 1: Bin Zhang Submission 3 , Team: Sasso ,
Link: https://www.synapse.org/#!Synapse:syn7199031

Model 2: Bin Zhang Submission 1, Team: ViResPred ,
Link: https://www.synapse.org/#!Synapse:syn7208407

Model 3: Bin Zhang submission 2 , Team: Aganita ,
Link: https://www.synapse.org/#!Synapse:syn7203047

We have performed 5 fold Cross Validation on the data set using the above 3 models to find out the best model. We have scored the model using the scoring.py. From there we selected the

model with the best score to create the Ensemble model leading us to the three models above. We did not have time to test them.
After creating the model there was imbalance in classification, which is why we used bagging algorithm while doing the Cross validation. This was given to us in the CV.R code. We got the predicted probabilities of the three models as output along with scores. We got predicted probabilities for all the 2371 CEL data and for each model. This gave us a matrix with 2371 rows which represent the CEL data and 4 columns - Target, KNN, Lasso and SVM. Here are the 10 predicted probabilities the first 10 patients for review.

CEL	Target	KNN	Lasso	SVM
1	1	0.571429	0.99236	0.896637
2	0	0	0.014061	0.2259
3	0	0.333333	0.142341	0.339816
4	0	0	0.002843	0.123896
5	1	1	0.982337	0.852658
6	1	0.5	0.814309	0.673931
7	0	0	0.153157	0.196794
8	0	0	0.000355	0.021214
9	1	1	0.994128	0.817432
10	1	0.4	0.999313	0.839466

Ensemble method :

To create the ensemble method we considered the predicted probabilities from the above matrix. We fitted two ensemble models : Average and SuperLearner.
Averaging :

Averaging is a type of Ensembling method. We found the mean of the predicted probabilities of the three models in case of regression problem or while predicting probabilities of classification problem. Here is a review of the output from the averaging method.

CEL	Averages
1	0.820142
2	0.079987
3	0.27183
4	0.042247
5	0.944998
6	0.662747
7	0.116651
8	0.00719
9	0.937187
10	0.74626

TABLE 1.2 AVERAGE ENSEMBLE METHOD

Super learner :

Super learner or stacking is the algorithm we used for creating the Ensemble model. The super learner algorithm is a loss based supervised algorithm that finds an optimal combination of predicted algorithms. Here a subset of variables are taken from the predicted probabilities and the algorithms in the library in Superlearner package is then fit into it. The output of the super learner is given below.

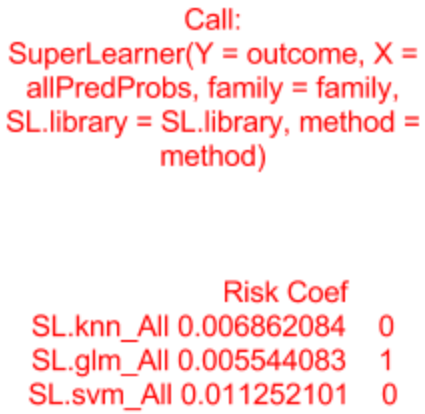

Figure 1.1 OUTPUT FROM SUPER LEARNER ALGORITHM

Technical Details and Learnings

An overview of the tools and softwares used in our program are given below.

	Software used	Tools
Data preprocessing	R	Sample, Cut
Feature selection	R	t.test
Model	R	KNN, SVM, Lasso
Scoring	python	Scoring.py
Ensemble	R	Average, Super Learner

Conclusion

We faced a lot of Challenges while doing the Project. Most of the challenge submissions and the Predicted probabilities from the Challenge submissions take a few hours to run the code and produce the Output. There was code compatibility challenges and learnings involved to understand and use the Challenge submissions and the Scoring.py code which was in Python. We learnt to use Python on Anaconda Navigator. We took During the coding and implementation, we faced system issues. The code was difficult to reproduce. There were some missing input files in the submissions. We were supposed to do a lot of Input output work for running the code. Some libraries used in the submission were difficult to find. We did not have time to test our Ensemble model because the code takes too long to print the output.

Future work :

We can improve our work by trying 10 fold cross validation instead of 5 fold to increase the accuracy. We can try using different Ensemble models. We can also try fisher's exact test so that we can access using contingency table.

References

1. Respiratory viral DREAM challenge 11: Ensemble of predictions. https://www.synapse.org/#!Synapse:syn5647810/wiki/399103
2. Challenge Submissions: https://docs.google.com/spreadsheets/d/1pN9-IIRNPO8lySeVYW1OuqWpLazfngPTtv_Zn4FJXvQ/edit#gid=0
3. "Wisdom of the Crowds": http://www.nature.com/nmeth/journal/v9/n8/abs/nmeth.2016.html (Links to an external site.)
4. Ensemble Methods: https://www.analyticsvidhya.com/blog/2017/02/introduction-to-ensembling-along-with-implementation-in-r/
5. We referred the previous year's submissions and Xiao's code and Final report for our project.

Division of Labor :

Presentation slide : Sonal Goswami

Code pre - processing :

Contributor : Aaron Devlin

Co-contributor : Amruthaa Rajan

Code PredictionprobAquire.r

Contributor : Aaron Devlin

Co - contributor : Amruthaa Rajan

Code Ensemble.R

Contributor : Amruthaa Rajan

Co-contributor : Aaron Devlin and Sonal Goswami

Research : Sonal Goswami

Documentation: *Project Report:*

***Read Me Document**:* Sonal Goswami