Predicting County Level Cost Differences for Treating Chronic Obstructive Pulmonary Disease

Jonathan Lin, Michael Smith, Aileen Wang
jolituba@stanford.edu, msmith11@stanford.edu, aileen15@stanford.edu

Motivation

Chronic Obstructive Pulmonary Disease (COPD) is the third leading cause of death in the US. The CDC estimates that COPD cost the US $32.1 billion in 2010, with 76% of the cost attributed to Medicare/Medicaid billings [1]. Our aim is to predict the increase/decrease in cost annually for a given county so that county and state officials can make more informed decisions about how to mitigate cost increases in their future.

Problem Definition

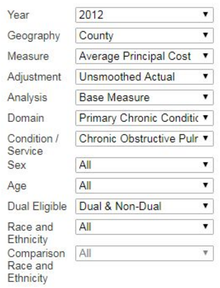
The goal of our project is to:
1. Find out the optimal feature set for predicting the cost
2. Binary Classification - Predict annual/multi-year cost increase or decrease for a given county
3. Multiclass Classification - Predict annual/multi-year cost increase or decrease percentage for a given county

Data Source

- Primarily gathered from Centers for Medicare and Medicaid Service database with features for each county
- Additional data from auxiliary sources
- Rows: county in a given year or a multi-year range
- Columns: features containing empirical data

Feature Selection

- Using the MMD tool below to download the csv file for each COPD measures:



- Merging all the csv file rows basing on the unique fips of each county
- Eight common features for single/multi-year prediction:
  - Urban or rural (boolean)
  - COPD Average Principal Cost (APC)
  - Emergency Department Visit Rate (EDVR)
  - Hospitalization Rate, COPD Prevalence
  - Prevention Quality Indicator (PQI)
  - Asthma Prevalence, Tobacco Prevalence
- Three additional features for multi-year prediction:
  - Preventive Services
  - Unemployment Rate (5y Avg.)
  - Percent Below Federal Poverty Level (5y Avg.)
- Predicted value: change in Average Principal Cost (APC)

Models

- Perform APC increase/decrease or increase/decrease percentage prediction using eight different models from Scikit-Learn Machine Learning package:

Model Name	SKlearn Method
Logistic Regression	LogisticRegression with 'multinomial' option is for multi-class prediction
Multi-layer Neural Network	MLPClassifier with a quasi-Newton solver='lbfgs'
K-Means Clustering	KNeighborsClassifier
Support vector machines (SVMs)	SVC with kernel mode ="rbf"
Decision Tree	DecisionTreeClassifier
Random Forest	RandomForestClassifier
AdaBoost classifier	AdaBoostClassifier - Adjusting the incorrectly classified instances for a given classifier
Naive Bayes	GaussianNB

- Pick optimal models by comparing prediction accuracy and number of optimal features

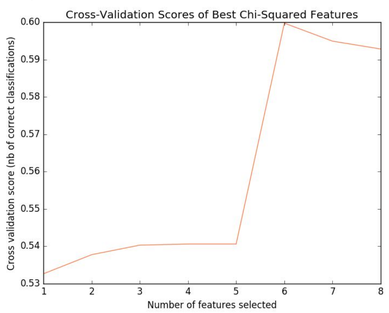
Results

- Input Data: 9498 rows for single year and 3168 rows for multi-year with 70% for training and 30% for test
- Baseline Result: Linear Least Squares:

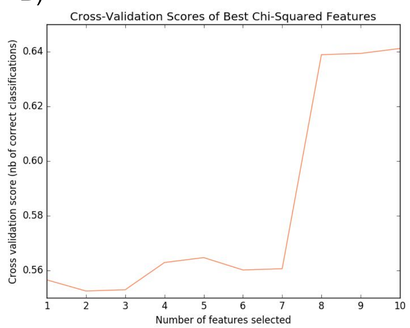
	Train Accuracy	Test Accuracy
Binary	60.5144%	59.4386%
Multiclass	33.6643%	34.2807%

- Optimal Features Selection Result:

A)

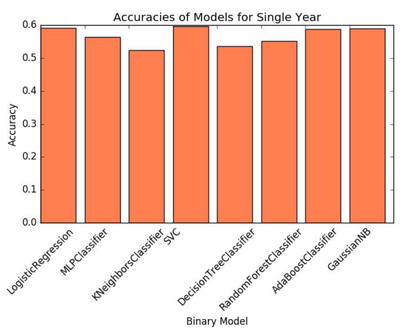


B)

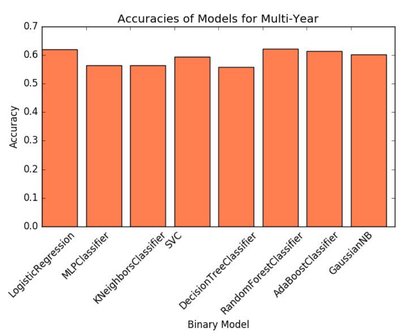


- Model Performance Comparison Result:

C)

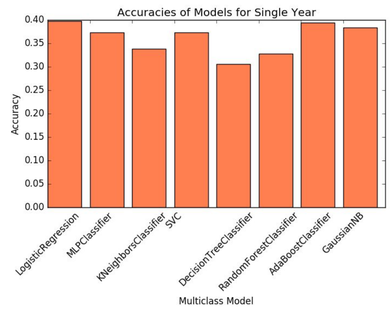


D)

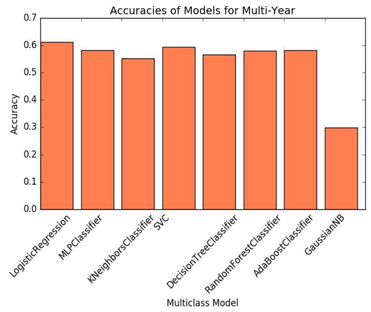


Results (cont.)

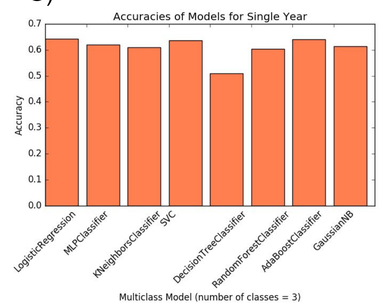
E)



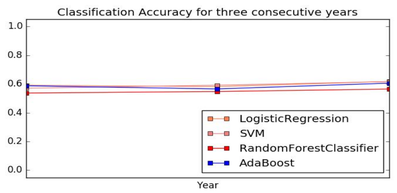
F)



G)



H)



Discussion

- We used Chi-squared as the scoring function to determine that the optimal number of features to be used was 6 for single year prediction and at least 10 for multi-year prediction (see Figure A and B)
- After running our binary and multiclass classifiers on the selected features, we observed that logistic regression, SVM, Random Forest, and AdaBoost are the top performers (See Figure C and D). Figure H shows the consistence of the accuracy
- For multi-year prediction, the above top four models performed well both for binary and multiclass classifications (see Figure D and F)
- For single year prediction, the multiclass classification performed more poorly than the binary classification (see Figure C and E). But, if reducing the number of classes to 3, then the multiclass classification performed well and even better the binary classification (see Figure G)
- We also tried to add more COPD related features, such as, percentage of forest cover, pollutants, benzene. But the data for these features were not available each year. It has degraded classification performance both for binary and multiclass.

Future

- Two possible issues: underfitting and/or issues with data
- To fix underfitting: use more detailed models (potentially with kernels and/or more features)
- To fix issues with data: ablative feature analysis and/or different features
- Add secondary classification predictions for more robust results

References

[1] CDC Features - Increase expected in medical care costs for COPD, Cdc.gov, https://www.cdc.gov/features/ds-copd-costs/index.html
[2] National Environmental Public Health Tracking Network Query Tool, Ephtracking.cdc.gov, https://ephtracking.cdc.gov/DataExplorer/\#/
[3] "Mapping Medicare Disparities", https://data.cms.gov/mapping-medicare-disparities , 2017