Geisinger Collider Project

Predicting COPD in pneumonia patients

Phase 2 - Spring 2016
Rebecca Barter and Shamindra Shrotiya
Question:

For a new patient who has been diagnosed with pneumonia, do they have Chronic Obstructive Pulmonary Disease (COPD)?

Can incorporation of external information improve prediction?
Chronic Obstructive Pulmonary Disease

- COPD is a major cause of mortality worldwide.
- Approximately 12 million adults in the U.S. having been diagnosed with COPD.
- A further 12 million adults in the U.S. are currently living with undiagnosed COPD.
Key Hypotheses: COPD Risk Factors

- **Smoking**
  - Available from Geisinger clinical data!

- **Occupational exposure to VOCs** (emissions from biomass fuels)
  - Infer from employment information provided by Geisinger!

- **Outdoor pollution**
  - Find on the internet

- **Weather**
  - Find on the internet
The Data Collection
Clinical Data

- **COPD:** patient is diagnosed with pneumonia at least once
- **Non-COPD:** patient is diagnosed with pneumonia exactly once, and not on the last visit.

State = Pennsylvania

- **COPD:** 18,567 individuals, 59,502 records
- **Non-COPD:** 9,622 individuals, 30,327 records

Age $\in [18, 90)$

- **COPD:** 19,038 individuals, 60,689 records
- **Non-COPD:** 9,867 individuals, 30,863 records

- COPD: select first record for each patient
- Non-COPD: select record corresponding to the single pneumonia visit

- **COPD:**
  - 5,704 individuals, 26,214 records

- **Non-COPD:**
  - 2,374 individuals, 10,346 records

- **COPD:**
  - 5,704 individuals, 5,704 records

- **Non-COPD:**
  - 2,374 individuals, 2,374 records
## Daily Summary Data

### Criteria Gases

<table>
<thead>
<tr>
<th>Year</th>
<th>Ozone (44201)</th>
<th>SO2 (42401)</th>
<th>CO (42101)</th>
<th>NO2 (42602)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>259,151 Rows</td>
<td>213,936 Rows</td>
<td>129,000 Rows</td>
<td>95,351 Rows</td>
</tr>
<tr>
<td></td>
<td>2,958 KB</td>
<td>2,198 KB</td>
<td>1,141 KB</td>
<td>1,309 KB</td>
</tr>
<tr>
<td></td>
<td>391,846 Rows</td>
<td>324,818 Rows</td>
<td>215,101 Rows</td>
<td>148,509 Rows</td>
</tr>
<tr>
<td></td>
<td>4,389 KB</td>
<td>3,277 KB</td>
<td>1,826 KB</td>
<td>1,991 KB</td>
</tr>
<tr>
<td></td>
<td>391,592 Rows</td>
<td>332,132 Rows</td>
<td>216,689 Rows</td>
<td>139,272 Rows</td>
</tr>
<tr>
<td></td>
<td>4,388 KB</td>
<td>3,340 KB</td>
<td>1,822 KB</td>
<td>1,841 KB</td>
</tr>
<tr>
<td></td>
<td>388,718 Rows</td>
<td>330,112 Rows</td>
<td>222,504 Rows</td>
<td>134,777 Rows</td>
</tr>
<tr>
<td></td>
<td>4,404 KB</td>
<td>3,335 KB</td>
<td>1,908 KB</td>
<td>1,776 KB</td>
</tr>
<tr>
<td></td>
<td>381,859 Rows</td>
<td>323,535 Rows</td>
<td>226,465 Rows</td>
<td>131,819 Rows</td>
</tr>
<tr>
<td></td>
<td>4,425 KB</td>
<td>3,241 KB</td>
<td>1,980 KB</td>
<td>1,770 KB</td>
</tr>
</tbody>
</table>
# Daily weather data from PSU Climatologist website

## PASC IDA Data Page

Select a network: FAA Daily
Select a display option: List Map

## Viewing Data Network FAA_DAILY

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>County</th>
<th>State</th>
<th>Lat</th>
<th>Lon</th>
<th>Elevation (feet)</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>KABE</td>
<td>ALLENTOWN</td>
<td>LEHIGH</td>
<td>PA</td>
<td>40.650</td>
<td>-75.440</td>
<td>376.0</td>
<td>1948-02-01</td>
<td>2016-04-07</td>
</tr>
<tr>
<td>KAOO</td>
<td>ALTOONA</td>
<td>BLAIR</td>
<td>PA</td>
<td>40.290</td>
<td>-78.320</td>
<td>1504.0</td>
<td>1977-01-28</td>
<td>2016-04-07</td>
</tr>
<tr>
<td>KBVI</td>
<td>BEAVER FALLS</td>
<td>BEAVER</td>
<td>PA</td>
<td>40.770</td>
<td>-80.390</td>
<td>1230.0</td>
<td>1996-01-02</td>
<td>2016-04-06</td>
</tr>
<tr>
<td>KBFD</td>
<td>BRADFORD</td>
<td>MCKEAN</td>
<td>PA</td>
<td>41.800</td>
<td>-78.640</td>
<td>2142.0</td>
<td>1957-07-01</td>
<td>2016-04-07</td>
</tr>
<tr>
<td>KBTP</td>
<td>BUTLER</td>
<td>BUTLER</td>
<td>PA</td>
<td>40.770</td>
<td>-79.950</td>
<td>1250.0</td>
<td>1992-02-26</td>
<td>2016-04-07</td>
</tr>
<tr>
<td>KCXY</td>
<td>CAPITAL CITY</td>
<td>YORK</td>
<td>PA</td>
<td>40.220</td>
<td>-76.850</td>
<td>340.0</td>
<td>0000-00-00</td>
<td>2016-04-07</td>
</tr>
<tr>
<td>KFIG</td>
<td>CLEARFIELD</td>
<td>CLEARFIELD</td>
<td>PA</td>
<td>41.040</td>
<td>-78.410</td>
<td>1516.0</td>
<td>2000-12-31</td>
<td>2016-04-07</td>
</tr>
<tr>
<td>KDLN</td>
<td>DOYLESTOWN</td>
<td>BUCKS</td>
<td>PA</td>
<td>40.330</td>
<td>-75.120</td>
<td>394.0</td>
<td>1999-07-28</td>
<td>2016-04-07</td>
</tr>
<tr>
<td>KDU1</td>
<td>DUBOIS</td>
<td>JEFFERSON</td>
<td>PA</td>
<td>41.180</td>
<td>-78.900</td>
<td>1814.0</td>
<td>1973-01-27</td>
<td>2016-04-07</td>
</tr>
<tr>
<td>KERI</td>
<td>ERIE</td>
<td>ERIE</td>
<td>PA</td>
<td>42.080</td>
<td>-80.170</td>
<td>730.0</td>
<td>1926-01-01</td>
<td>2016-04-07</td>
</tr>
</tbody>
</table>
Sounds great!

So what’s the problem?
The smoking and employment information was missing!
The EPA “daily” values were not daily at all...

[Bar chart showing the percentage of values for different variables (CO, NO2, Ozone, SO2, PM10, PM2.5, Arsenic, Lead, NO, CS2). The bars are divided into green (not missing) and grey (missing) sections.]
We went ahead and blended anyway...
Data Blending

- Blended EPA and PSU with Geisinger using **date** and **closest zipcode**.
- **Problem**: not a lot of geographical overlap between Geisinger patients and environmental measurement sites.
Pre-processing the data before modeling:

Dealing with missing values
How can we deal with missing values?

Possible ideas:

1. **Exclude all** observations that had any **missing features** to only leave a modeling dataset with no missing data.
2. Utilize methods that **directly allow for missing data** in the modeling process.
3. **Exclude** all features that have more than a **threshold proportion** of missing values.
4. Perform **imputation** on all missing features using the **median**, mean or a k-nearest-neighbors approach from non-missing values from the same feature.
How did we deal with missing values?

**Our approach:** to minimise data loss and ensure practicality

- **Remove** all variables with more than 8% missing values.
- **Impute** the remaining missing values
  - Numerical features: impute using the median.
  - Categorical features: impute using the mode.
Pre-processing the data before modeling:
Dealing with unbalanced classes
How can we deal unbalanced classes?

- Many machine learning algorithms are known to perform poorly under class imbalance.
  - We have 5,704 COPD patients and 2,374 non-COPD patients.

Possible ideas:

- **Upsample**: randomly sample labels from the smaller class (non-COPD) with replacement to be equal in number to the non-COPD labels.
- **Downsample**: randomly sample labels from the larger class (COPD) to be equal in number to the non-COPD labels.
How did we deal unbalanced classes?

Our approach

- **Upsample**: randomly sample labels from the smaller class (non-COPD) with replacement to be equal in number to the non-COPD labels.

No need to sacrifice sample size.
Stepwise Feature Inclusion
Stepwise Feature Inclusion

- **Geisinger Clinical**
  - Gender, marital status, employment status, age, race, asthma

- **Geisinger Clinical + Smoking**
  - Gender, marital status, employment status, age, race, asthma
  - *binary smoking variable*

- **Geisinger Clinical + Smoking + PSU weather data**
  - Gender, marital status, employment status, age, race, asthma
  - binary smoking variable
  - *average temperature, pressure and humidity in the week preceding the admission*
Modeling
We used empirically well-tested non-parametric models

- Random Forest
- GBM
- XGBoost

To fit these models we used the R **caret** package

- Test interaction of various combination of input parameters e.g. for GBM varied interaction depth and number of trees
- Parameters selected using 5 repeated rounds of 10-fold CV
Results
Results

- **Black diamond:** average of CV estimates for the optimal parameter set.
- **Red circle:** prediction accuracy on withheld test set.

**Best performing model:**
- Random Forest
- **Accuracy** of 70%
Conclusion
Conclusion

**Question:** For a new patient who has been diagnosed with pneumonia, do they have COPD?

- Data collected was plagued by missing values.
- Better performance accuracy may have been achievable with better quality data:
  - complete *smoking pack-years*
  - outdoor and indoor *pollution data*