Motivation

- The US spends $332 billion on diabetic and prediabetic care every year.
- Diabetes affects 350 million people with 3 million dying each year due to complications.
- Through analyzing diabetic patient data, we hope to help reduce the mortality rate of diabetic people through improving patient care and decrease its cost.
- We will analyze data from 130 hospitals in the United States from 1999 to 2008 to create 2 models that will:
  1. Predict whether a diabetic patient will be readmitted to the hospital in less than 30 days (i.e., a binary model).
  2. Predict the probability of readmission within 30 days.
- Doctors can use these models during patient visits to guide patient care decisions.
- The binary model can be used to infer general patterns in the data and for holistic research on early hospital readmittance.

Method

- We decided to pursue a variety of machine learning techniques and assess their performances to find our optimal model.
- We looked at the coefficients for our logistic regression models for some preliminary inference analysis. Overall, we found:
  - Early readmittance is negatively correlated with seeing doctors with the specialties of Immunology, Sports Medicine, Pediatrics, Surgery, Colon & Rectal Surgery.
  - Early readmittance is highly correlated with overall readmittance.
  - Early readmittance is highly negatively correlated with the increased dosage of Chlorpropamide & Giburide Metformin.

Analysis and Results

Preprocessing:
- When running our experiments, we discovered 7 columns that provided little predictive value.
- They were too sparsely populated and were eliminated from the dataset.
- We looked at the coefficients for our logistic regression models for some preliminary inference analysis. Overall, we found:
  - Early readmittance is negatively correlated with seeing doctors with the specialties of Immunology, Sports Medicine, Pediatrics, Surgery, Colon & Rectal Surgery.
  - Early readmittance is highly correlated with overall readmittance.
  - Early readmittance is highly negatively correlated with the increased dosage of Chlorpropamide & Giburide Metformin.

Analysis of Models:
- Established 0-1 loss as standard metric by which we would assess accuracy.
- Use 0-1 loss of naive classifier, which would predict the most frequently occurring label (in our case, 0), as our baseline for comparing accuracies.

For greater insight into their performance, we found our models’ ROC curves and AUC values. The ROC curves of our 3 most accurate models are included below:

<table>
<thead>
<tr>
<th>Model</th>
<th>Model Description</th>
<th>Train AUC</th>
<th>Test AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Classifier</td>
<td>Predict most common label (in our case, 0)</td>
<td>0.810754</td>
<td>0.809576</td>
</tr>
<tr>
<td>Binomial Log Reg</td>
<td>Used features (binary output)</td>
<td>0.810286</td>
<td>0.809752</td>
</tr>
<tr>
<td>Binomial Log Reg</td>
<td>Used features determined to be significant (as in original study)</td>
<td>0.811032</td>
<td>0.809970</td>
</tr>
<tr>
<td>Multinomial Log Reg</td>
<td>Used features (binary output)</td>
<td>0.811029</td>
<td>0.809703</td>
</tr>
<tr>
<td>Elastic Net</td>
<td>Log reg with both L1 and L2 regularization, used all features</td>
<td>0.810459</td>
<td>0.809837</td>
</tr>
<tr>
<td>Random Forests</td>
<td>Use of bagging and bootstrapping to generate set of decision trees</td>
<td>0.81352854</td>
<td>0.80984266</td>
</tr>
<tr>
<td>SVM</td>
<td>SVM solver, rbf kernel</td>
<td>0.8125</td>
<td>0.8077</td>
</tr>
<tr>
<td>SVM</td>
<td>SVM solver, lin. rbf kernel</td>
<td>0.8165</td>
<td>0.8092</td>
</tr>
<tr>
<td>SVM</td>
<td>SVM solver, gaus. rbf kernel</td>
<td>0.8125</td>
<td>0.8077</td>
</tr>
</tbody>
</table>

For greater insight into their performance, we found our models’ ROC curves and AUC values. The ROC curves of our 3 most accurate models are included below:

Future Areas for Investigation

- On top of our existing models, we aim to create a classifier featuring a neural network.
- With this neural network, we will experiment with different activation functions.
- In addition, we will do additional analysis on the results of these models.

Conclusions

Weight analysis:
- Consulting doctors with particular specialties (e.g., pediatrics, sports medicine, dermatology, colon & rectal surgery) are very negatively correlated with readmittance.
- For diabetics, either the doctors in their specialties frequently deal with cases strongly related to diabetes or cases that are orthogonal to diabetes.
- Having been readmitted is a strong hint towards future readmittance.
- Hints that if a case’s root cause has been resolved, there is low likelihood of readmittance.
- Drug prescriptions indicate the relative severity of the problem.
- E.g., an upwards dosage of chlorpropamide has a very negative weight; since chlorpropamide only deals with minor diabetic issues, this would signal the the case is most likely not very severe.

- Analysis of correlation trends can be very valuable to both doctors and hospitals.
- These models can help doctors make patient care decisions.
- E.g., the strong correlation between past readmittance and early readmittance could help doctors be particularly vigilant on readmitted patients.
- Models for prediction can help hospitals with scheduling as well as help insurers more effectively gauge risk.

Acknowledgements

Thank you to Lucio Dery, Professor Boneh and Professor Ng for their guidance and support!

Citations

1. “Beating Diabetes: Predicting Early Diabetic Patient Hospital Readmittance to Help Optimize Patient Care.” Charlie Xu, Christina Pan, Stephone Christian. {cxu2, capan, stephone}@stanford.edu