Heritage Health Prize Competition

I. Competition Synopsis

The Heritage Health Prize Competition occurred over the duration of two years. The primary goal was for a team to develop a winning algorithm capable of utilizing electronic health records in order to accurately predict how many days a given patient would spend in the hospital in the next year, ideally revolutionizing the health care system by significantly decreasing unnecessary hospital admissions.

The *Market Makers (MM)* team won prizes for both Round One and Two of the Kaggle competition. Their winning methods elucidated key data processing and analysis techniques. *MM* first aggregated data to the patient level and split the resulting information into two categories; the first model consisted of year- long data. If a patient had claims over two years, each year was considered a unique record. The second model aggregated patients by year and removed data in which there was not a two year record. This resulted in greater claim history per patient but less data overall, and thus less resolution.

The goal of *MM*'s predictive modelling was to minimize the RMSE (root-mean-square error) between predicted and actual days spent in the hospital for their given testing data. *MM* originally utilized four primary algorithms and determined Gradient Boosting Machines to be the most effective. The primary concern of the group was over- or under- fitting data. As a result, *MM* utilized a technique of blending various prediction techniques to account for flaws in each other, arguing that "different algorithms can arrive at their solutions by different paths – they look at the problem from different perspectives. When these algorithms are combined there is resulting synergy" (Brierley 5). *MM'*s final model utilized 79 other models. They used a combination of algorithm dependent and independent predictions, a method that proved to be successful in a previous Kaggle competition. Their algorithm dependent prediction model utilized linear regression- specifically ridge regression to address extreme weights- to assign weights to various models, thus utilizing each one to a different level depending on their resulting accuracy. In contrast, the independent prediction model simply took the median of models built using varying data sets.

Looking into *MM* and other teams' approaches, we would take a similar approach to the problem, but aggregate on a single year and employ more modern machine learning approaches, such as neural networks used by natural language processing experts, to learn patient behavior over time. We would also try to associate ClaimID with LabCount and DrugCount entries so that a mapping to claims and lab/drug history can be established rather than taking an aggregate count.

II. Data Analysis

Question

Inspired by the Heritage Health Prize Competition results, we were curious to see if electronic health record data could be further extrapolated to predict which department a patient would most likely be

admitted to. In theory, the ability to predict a patient's trajectory within the hospital could reduce patient wait time and better manage traffic flow-- a crucial variable for patients in critical conditions.

Feature Extraction

The first step to any data analysis project is choosing appropriate features for the model. In our case, we decided to build a supervised machine learning model, where we would train the model on the chosen features and correlate it to the procedure group for each patient's claim (y variable). We referred to the Data Dictionary provided to choose appropriate features for our model. After thorough conversation and trial/error, we decided to use data from the Claims, Members, LabCount, and DrugCount datasets to extract features for our model. We used MemberID, Year, PlaceSvc, and Procedure Group from the Claims dataset; MemberID, AgeAtFirstClaim, and Sex from the Members dataset; MemberID, Year, and LabCount from the LabCount dataset; MemberID, Year, and DrugCount for the DrugCount dataset.

We chose these specific features because MemberID and Year are indicative of a patient's health history in a given time span (a year). We used PlaceSvc, because the place of service elucidates how the patient was brought into care. If a patient was brought in via an ambulance, independent lab, or office, this will likely have an impact on the final Procedure Group. We also decided to use AgeAtFirstClaim and Sex, because these features may correlate with specific procedures, such as cardiovascular surgery for males. We also looked at the DrugCount and LabCount of the patients because we thought that the number of lab tests and drugs prescribed may be indicative of the severity of the procedure. For example, if a patient is admitted into the hospital by an ambulance and has a history of several lab reports and drugs administered within the past year, that may indicate a high likelihood that surgery is necessary.

Data Processing

With our features decided, we adopted techniques inspired by winning papers from the Heritage Health Foundation to pre-process our data. Because the entries in the Claims dataset are disjoint from the entries in the LabCount and DrugCount datasets, meaning there lacks a one-to-one mapping between the entries in the datasets, we employed a technique that *MM* used in their Milestone 1 and 3 reports; we aggregated the Claims, LabCount, and DrugCount datasets based on Year and MemberID. This ensured that each entry mapped to a member in a specific year and provided an accurate mapping between entries in the Claims dataset vs. LabCount and DrugCount datasets. We chose a one year frame, because we thought that a single year would be an appropriate indication of a patient's health and is also the standard for insurance policies given to us in the dataset. Ideally, we would like to extend the depth of our study via monthly-specific time frames.

Next, we had to convert string representations of values into numbers. We converted the AgeAtFirst, LabCount, and DrugCount columns into integers and removed "+" characters presented in some entries. After grouping the LabCount and DrugCount columns by MemberID and Year and converting the data into the proper integer type, we also took the sum of the Lab/DrugCount columns per year for each member and left joined it with the Claims dataset, which was also grouped by member and year. This ensured that all members who had lab examinations or drugs administered within that year were given the same lab and drug count. For those who did not have any lab or drug counts in a particular year, their lab and drug count was set to 0. We

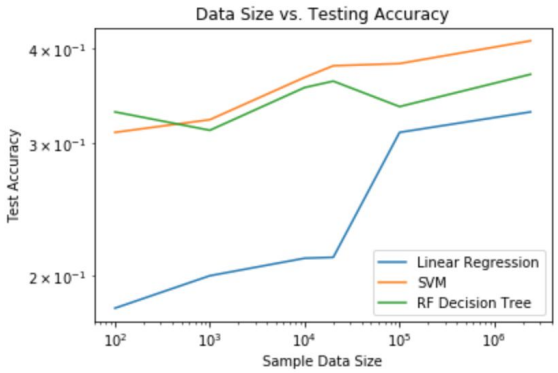
AgeAtFirstClaim	AgeApproximation
0-9	5
10-19	15
20-29	25
30-39	35

dropped records without ProcedureGroup, Age, Sex, and PlaceSvc, because those were considered essential variables. ProcedureGroup was also our question's label, so it could not be Null.

Finally, we had to convert categorical variables into numeric, binary representations. This is important, because most machine learning models cannot classify on string representations of variables. Thus, we one hot encoded PlaceSvc and Sex to be binary representations of each other, and dropped one column from each OHE, the Sex_F and Place_Svc_Other columns, to ensure that there were not any linear dependencies between features.

Results and Conclusion

We tried various sklearn models in Python to build our model. We tried multinomial Logistic Regression, SVMs, and Random Forest Decision Trees to classify our data, with k-fold validation and hyper-parameter tuning for max_depth and n_estimators. Given the limited computing power, size of the dataset, and time to execute this project, a multinomial SVM showed to be the best classifier, allowing us to achieve up to 41% in test accuracy and 99% in training accuracy. This means we are likely overfitting to the training data and might need to change our features to make them more indicative of the data. If we had more time, we would use gradient boosting machines, regression trees with more hyper-parameter tuning, or neural networks on a cluster to do more in-depth analysis.



Overall, this project gave us valuable insight on how difficult extracting and cleaning EHR data is. Even though this dataset was relatively "clean," we still had to employ various data pre-processing techniques to ensure that our features, results, and data were meaningful and applicable in a real medical setting. We also realized how the gaps between EHR data and medicine in practice are very different. For example, we weren't given ClaimID's, the patient's full medical background and history (past illnesses/ injuries), and even current age at the time of the claim. Additional data points including but not limited to those mentioned above in conjunction with expert advice from healthcare professionals could help us extract more powerful features and build a more comprehensive model.

References

Brierley, Phil, David Vogel, and Randy Axelrod. "Heritage Provider Network Health Prize Round 1 Milestone Prize: How we did it–Team 'Market Makers'." (2011).

"Heritage Health Prize." Kaggle, www.kaggle.com/c/hhp.