Hospital length of stay is an important measure that has an impact on many things including costs to the patient and facility, availability of services for others, and even recovery time for the patient. Understanding factors that impact length of stay can help facilities better plan for bed availability and resource needs, and potentially find areas that can help shorten length of stay for the benefit of the patient. As the primary provider of hospital services in the Wausau area, it’s important for Aspirus Hospitals to be able to predict and potentially influence length of stay for their own benefit and for the benefit of patients that rely on the for care.

Using the *Health Care Analytics – 2* data set obtained from Kaggle (<https://www.kaggle.com/vetrirah/av-healthcare2>) provides data that includes several factors including Hospital Type Code, Hospital Region Code, Available Extra Rooms in Hospital, Department, Ward Type, Ward Facility Code, Bed Grade, Admission Type, Severity of Illness, Number of Visitors, Age, Admission Deposit, and Length of Stay. After excluding 113 observations that were missing Bed Grade, there are 313,906 observations that include Length of Stay in this data set, providing a substantial amount of data to use as a baseline for creating a model. These variables are all potential impacts to how long a person stays in the hospital, particularly when combined with one another.

The data, both the potential predictors and the response variables, are primarily categorical data. This lends itself well to analysis with Association Rules, which can be useful in finding patterns in the data as well as what combination of variables have the most impact on length of stay. One of the factors is the number of visitors with a patient, which I considered excluding since it isn’t something that is known at the time of admission. I chose to keep it as part of the model for two reasons. First, length of stay isn’t something that should be reviewed only upon admission, but also on an ongoing basis when this data can be considered. And second, I think this is an important consideration and may also reveal more insight as more data is gathered from hospital stays during COVID-19, when visitors are restricted. I also converted the length of stay categories into two categories, differentiating between stays that were over and under 30 days, in order to be able to better understand the impact combinations of predictors had on length of stay.

I was also interested in analyzing how the factors that contributed to the top 10 association rules, based on lift, would contribute to a logistic regression model. To make this connection, I chose the predictors that were included in more than one top 10 rule as the predictors in a logistic regression model. I compared this to the full model (excluding factors with high collinearity) and to a model that included predictors that were more commonly thought of as influencing length of stay (Admission Type, Severity of Illness, Department, and Age).

The top 10 Association Rules with factors that led to a length of stay over 30 days is shown in figure 1. The lift and confidence are also included. While there are some commonalities among the rules, the differences highlight that different combinations of predictors can be used to together to determine if a length of stay will be longer. By considering the variables for a single admission, the likelihood of the stay being more or less than 30 days can be determined using this model. For example, a patient that is admitted to Ward Type R, with and Admission Type of Trauma, with 4-32 visitors (the maximum number), and in the Gynecology Department is 2.43 times as likely to have a stay of over 30 days. Comparatively, a patient that is admitted to Ward Type R, with and Admission Type of Trauma, with 4-32 visitors (the maximum number), but NOT in the Gynecology Department has a slightly lower lift, being 2.33 times as likely to have a stay of over 30 days. This demonstrates where slight changes in the admission can have an impact on the potential for a longer length of stay and can help the Hospital plan more effectively for care and bed availability.

The logistic regression model comparison indicates that using all non collinear predictors (which excludes Hospital Type Code, Ward Facility Code, and Hospital Region Code) leads to a slightly better model than using the predictors found in the top 10 association rules, and a much better model than using more common predictors. This was demonstrated in the AIC, the AUC of an ROC Curve, and the error rate from cross validation. The full model had the lowest AIC, lowest CV error rate, and highest AUC (0.7859) among the three models. The ROC Curves and resulting AUC is shown in figures 2,3 and 4. I think this in turn supports using all of the predictors in the association rules, even those that may seem less obviously related to the length of stay. These models together can be beneficial in providing Aspirus with insights it needs to be able to plan effectively for care of its patients.

Figure 1

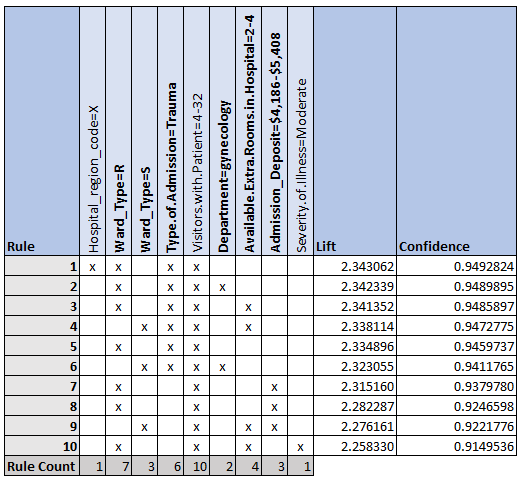


Figure 2

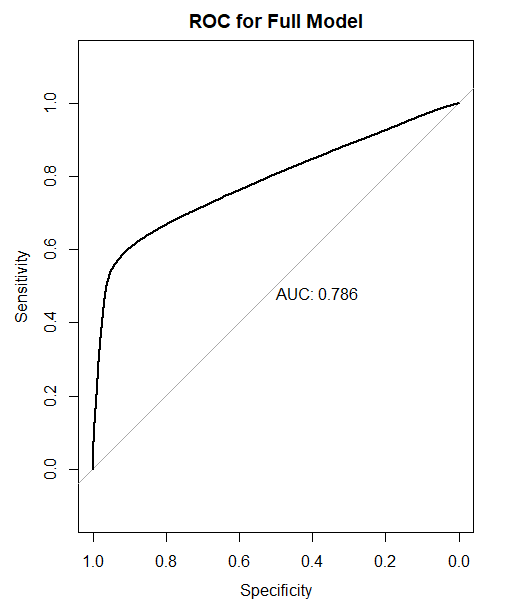


Figure 3 Figure 4

