# Predicting and Reducing Diabetes-Related Readmission Risk: A Big Data Analytics Approach

J. W. and Alan Chihwaro

San Diego State University

# Executive Summary

In 2011, the estimated annual cost attributed to unplanned hospital readmissions in the United States reached $41.8 billion (Hines, Barrett, Jiang, & Steiner, 2014). The primary aim for our project was to reduce the significant monetary burden that these readmissions place on the national healthcare system. In pursuit of this objective, we set a goal of creating a predictive model that is able to determine, with 65% accuracy or better, which patients will be readmitted after a medical encounter. To generate this model we used a dataset sourced from the UCI Machine Learning Repository that covers 10 years of patient data from 130 hospitals. Having data from a wide range of hospitals was particularly desirable because we wanted our model to be as widely applicable as possible. The raw dataset was composed of 50 variables detailing information on over 100,000 diabetes-related patient encounters. In preparing the dataset for the final model we conducted a number of data preparation procedures including: removing variables due to large numbers of missing variables, removing highly correlated variables, removing variables with very low variances, removing records with missing values, and recoding variables with many factor levels. Our final, processed dataset was left with 27 variables providing the details of 80,631 patient encounters.

Logistic Regression, Decision Tree, and Random Forest modeling was conducted to identify significant variables in predicting whether a diabetes-related inpatient encounter resulted in patient readmission. We used a variety of modeling techniques to validate and improve our model performance including: cross-validation, sampling for balancing class distributions, and the use of a training and test set. Of the three different models we built, our random forest model had the highest overall accuracy, 68.19%, and the highest accuracy in predicting readmission when the patient was actually readmitted, 64.54%. The most significant predictor in all three of these models was the number of prior inpatient visits. Interestingly, the risk of readmission was found to decrease for those with a higher number of prior inpatient visits. We suspect this result is due to the fact that doctors already identify prior inpatient visits as a readmission risk and tend to monitor those patients more closely. The number of diagnoses entered into the system was the second most significant predictor of readmission risk in that patients with greater than 6.5 diagnoses are likely to be readmitted. Additionally, our results indicate that the likelihood of readmission increases if a patient was admitted to the hospital from another care center or had a primary diagnosis categorized as other. Alternatively, the likelihood of readmission decreases for patients who had a higher number of inpatient visits within a year prior to an inpatient encounter or who had a greater number of lab procedures conducted during their inpatient encounter. We also found all socio-demographic factors significantly correlated to readmission risk. Older patients were found to be much more likely to be readmitted than younger patients and Caucasian females were found as the most likely to be readmitted.

The success of our preliminary efforts warrants continued funding for the development of readmission risk predictive models. Firstly, our current model should be implemented as a pilot program in a handful of hospitals. Lessons learned from this pilot program can help improve our model before more widespread deployment. In concert with this pilot program, funding should also be provided for the development of universal healthcare data collection and storage technologies. Improved data collection methods can help reduce missing data caused by both technological and procedural errors. The combination of more robust datasets as well as real-world experience in running our model will lead to a successful rollout across a large number of healthcare facilities.

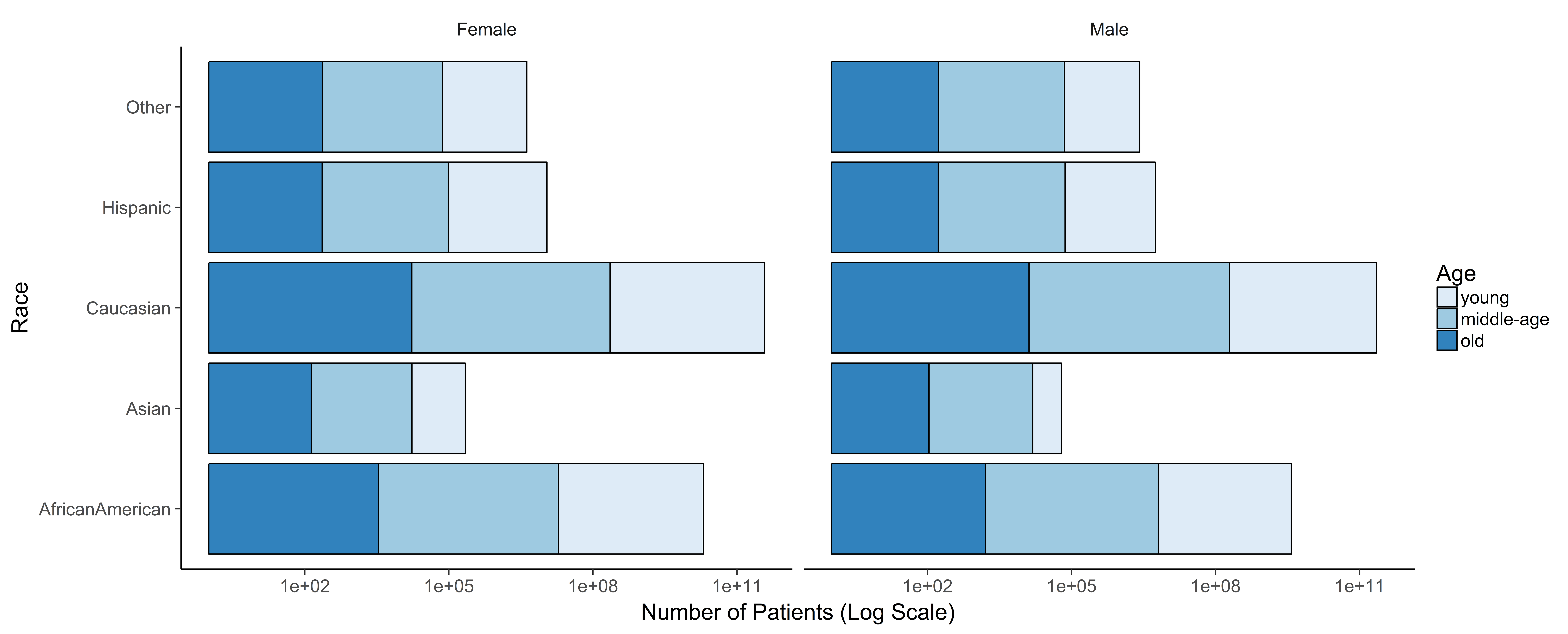
# Discovery

Since the high-level goal for this project was to reduce healthcare costs on a national level through the prediction of patient readmission, we were particularly interested in datasets that would give us access to information from more than just one hospital or region. We eventually found a suitable dataset from the UCI Machine Learning Repository that covered 10 years of patient data from 130 hospitals across the United States. The original dataset included 50 variables covering a wide range of information on over 100,000 patient encounters where some form of diabetes was diagnosed. Since diabetes is one of the leading conditions that causes readmission amongst Medicaid patients (Hines et al., 2014), we felt that developing a predictive model for diabetes-related readmission would lead to significant cost reductions as well as quality improvements.

After processing, our final dataset includes 27 variables on 80,631 inpatient visits. Initially imported from a CSV file, our dataset was stored and processed in R as a data frame throughout our analysis and development. 19 of the 27 variables in the dataset record categorical information about patient visits. These include things like demographic information, diagnoses, and readmission status. The remaining eight variables are all numerical data about either patient history or the encounter in question. These variables include the time spent in hospital, the number of various procedures performed, and the number of medications the patient is taking.

Demographics information in the dataset was recorded through three variables: race, gender, and age. Most notable about the distribution of patient demographics is the extremely large number of Caucasian patients compared to other races. For example, the dataset includes 61,179 observations of Caucasian patients while there are only 530 observations for the least represented race, Asian. Since there is such a large discrepancy in patient race, we used a logarithmic scale in Figure 1 below to more clearly show the distribution of the three demographic variables.

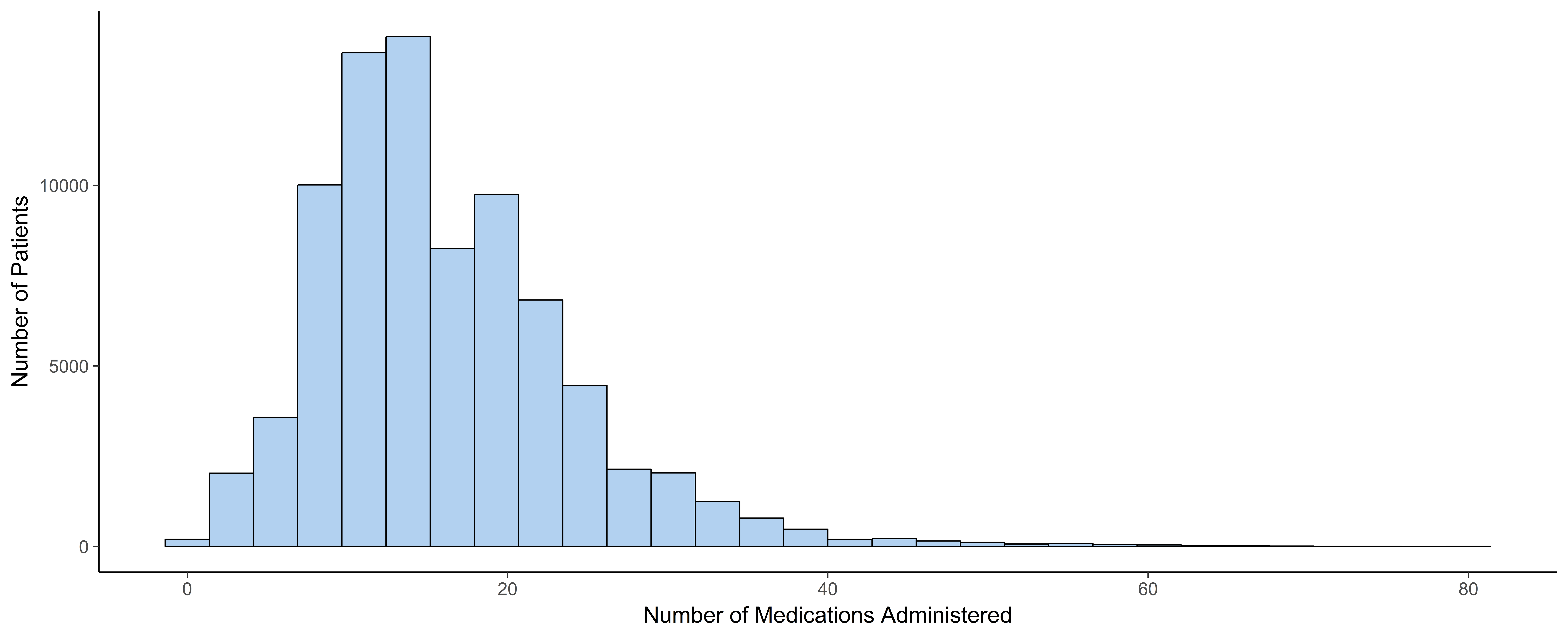
The data set also tracks admission and discharge information via three variables: admission type, admission source, and discharge disposition. Admission type is divided amongst patients coming from emergency or trauma centers, urgent care, and elective or newborn procedures. Admission sources include referrals, transfers from other care centers, and emergency rooms. Patient discharges were categorized as discharged to home, admitted to other healthcare provider, and discharged to home with care.

**

*Figure 1.* Faceted bar chartshowing distribution of race, age, and gender in the final dataset.

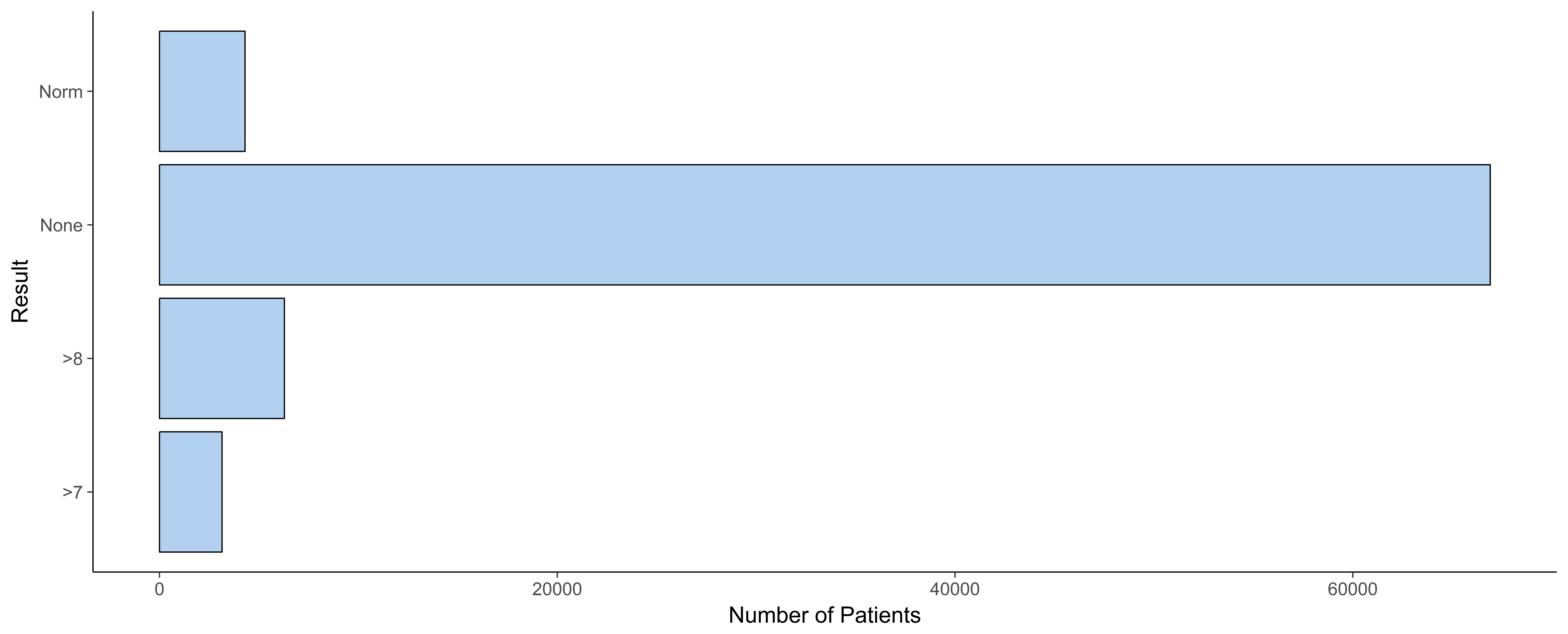
Details of medical and laboratory procedures during the encounter as well as the year preceding the encounter are also tracked in the dataset. The variables recording this information include: days in hospital, number of lab procedures, number of non-lab procedures, number of medications administered, number of inpatient visits, number of outpatient visits, number of emergency visits, and number of diagnoses. This aspect of the dataset provides information regarding current and past patient condition which is crucial when building a model to predict future condition. All of the numeric variables are included in this chunk of the dataset and they are almost all heavily skewed. For example, Figure 2 below displays the frequency of medication count.

In addition to tracking the number of diagnoses, the dataset also includes three variables recording primary, secondary, and additional secondary diagnoses for each encounter. Each of these variables is split into the following general categories: circulatory, respiratory, digestive, diabetes, injury, musculoskeletal, genitourinary, and other. Interestingly, circulatory and other are the most common diagnoses for all three of these variables.

**

*Figure 2.* Histogram showing distribution of the medications administered variable.

The next general category of variables reflects data on medications and specific blood tests. There are six variables that indicate any recorded changes to dosages of the following medications: metformin, glipizide, glyburide, pioglitazone, rosiglitazone, and insulin. Additionally, there is one overall variable that indicates whether there were any doctor-prescribed changes to any medication dosages and another that indicates whether a new diabetes medication was prescribed. The A1Cresult variable records the results of the A1C test, which is a blood test that reflects blood sugar level and is used to diagnose diabetes. Interestingly, as seen in Figure 3 below, in the vast majority of cases an A1C test was not conducted.



*Figure 3.* Bar chartshowing distribution of A1C results.

The final remaining variable in the dataset is our outcome variable. It records whether the patient was readmitted following the encounter or not. In total, 38,385 patients were readmitted while 42,246 patients were not. The model planning and model building sections below include an in-depth discussion on the imbalance in the classes.

# Data Preparation

All of our data preparation was conducted within R using standard functions as well as the car package for its recode function. We started preparing our data for model building by pruning variables for a number of reasons. The original dataset contained multiple variables with a high proportion of missing data. These included weight (97% of values missing), medical specialty (47% missing), and payer code (40% missing). The weight variable was removed from our dataset because the sparsity of the data was too significant for the data to provide value to our analysis. Medical specialty was also excluded from our analysis due to the high-proportion of missing data and its minimal value contribution to our problem analysis. While medical specialty of the admitting physician could be an influential factor in determining risk of readmission, it would be too difficult for hospital management to control in improvement efforts. Payer code was considered relevant to readmission risk, but was removed due to the fact that relationship between payer code and readmission is likely to be captured by other factors of our dataset. For example, the number of procedures done, the nature of the patient’s discharge, and the time in hospital may be indirectly related to payer code in that the amount and type of services covered by the patients’ providers varies with each provider. We next identified and removed all the variables in our dataset with near zero-variance. With the exception of the glucose serum test result, all other near zero variance variables measured the prescription status or dosage change of various diabetes medications among the patients.

Of the remaining variables, eight of them had some form of missing values. Either the values were missing altogether or values such as “not available” or “null” were used to represent missing data. In preparing our data for analysis we removed any records with missing values from our dataset. All observations with a death-related discharge disposition were also removed from our dataset since they were considered inapplicable to predicting readmission. Additionally, any patient encounter whose discharge disposition indicated that they had left against medical advice was removed since they could potentially bias our results.

Age for the patients included in our dataset was originally grouped into intervals of 10. The proportion of patients in each age group was widely dispersed and highly uneven. We transformed the age attribute into larger intervals chosen based on the relationship between age and diabetes. According to the Centers for Disease Control and Prevention [CDC] (2015), type 1 diabetes onset in uncommon after age 40 and the average age for type 2 diabetes diagnosis is between 45 and 65. Based on these relationships, we grouped age into the following intervals 0 to 40, 40 to 70, and 70 to 100.

The dataset also contained multiple attributes where the number of patients within each factor level varied significantly, the counts were extremely small, or there were a large number of factor levels. While various analysis techniques can handle factors with many levels, the differences between the maximum and minimum values for the admission type, admission source, discharge disposition, and diagnoses were great enough that they could greatly bias our results. Additionally, many of the levels were interrelated so we used clinical domain knowledge to group related values for these variables together*.* This data transformation helped minimize the large variation in the number of observations of each possible value for these attributes as well as the potential bias that could be introduced into our model due to the fact that some factor levels contained an extremely small proportion of our population.

The final step in our data preparation was to consolidate our outcome variable, readmission status, into a binary class. The details of that change and how it impacts model building is discussed in the following section.

# Model Planning

The original dataset had recorded the readmission status of patients using a multinomial variable that identified patients as having been readmitted either within 30 days of discharge, after 30 days from discharge, or not at all. For our analysis we chose to look at whether or not a patient was readmitted instead. Therefore, we created a new variable, readmitted binary, for our final dataset consisting of 80,631 inpatient visit observations with which to conduct our data modeling. This classified the diabetes-related inpatient encounter as either having resulted in the patient being readmitted or not readmitted. As shown in Figure 4 below, this binary readmission outcome variable had 38,385 observations in the unplanned readmission class and 42,246 observations in the not readmitted class.

*Figure 4.* Class distribution of final diabetes dataset.

This uneven class distribution has the potential to bias our results in favor of the majority class, not readmitted. Therefore, during our model building we plan to deal with this class imbalance using either up-sampling, down-sampling, or a hybrid sampling approach to ensure an even split of our classes. Using this new outcome variable we were able to test our initial hypotheses regarding the relationship between readmission and a variety of factors describing inpatient encounters. These initial hypotheses are identified below.

**Initial Hypotheses**

H1: Our first hypothesis stated that a significant correlation exists between patient socio-demographics and patient readmission.

H2: Our second hypothesis was that the number and type of medical encounters preceding the recorded visit are significantly correlated with readmission risk.

H3: With our third hypothesis we predicted that the number of inpatient lab tests of diabetes patients has a significant impact on risk of unplanned readmission.

H4: Our fourth hypothesis predicted a significant negative relationship between readmission and a patient’s insulin levels. Specifically, we hypothesized that steady insulin levels for people with diabetes-related inpatient visits indicates a lower likelihood of an unplanned readmission. We chose to conduct a logistic regression analysis with insulin levels as a predictor of unplanned readmission to test this hypothesis.

H5: Our fifth hypothesis changes in diabetes medication prescribed as a result of an inpatient visit indicate an increased risk of readmission.

To evaluate our first hypothesis we selected to use a chi-square test to determine whether a relationship exists between readmission and the race and age of the patient. We chose to use logistic regression modeling to identify whether or not the variables in our second to fifth hypotheses were significant in predicting readmission. In addition to our individual hypothesis testing, we planned to build a comprehensive model for identifying the key predictors of unplanned readmission. A mix of 27 categorical and numerical descriptor variables on inpatient observations were selected to use for training this model. 17 of these descriptor variables were factors of patient socio-demographics, medical history, and health outcomes and 10 were attributes on the clinical care process. The outcome we are trying to predict is whether or not a diabetes-related inpatient encounter resulted in a readmission. Since this is a binary outcome, readmitted or not readmitted, we selected to use logistic regression, decision tree, and random forest modeling to identify the most influential factors for predicting unplanned readmission. To facilitate evaluation of how accurately these influential factors can predict readmission, we will use a training and testing set for model building and model performance evaluation.

# Model Building

In building our model, we initially investigated the appropriateness of imputing the missing values in our data. We used margin plots and parallel box plots to observe the distributions of missing values and determined that the values were not missing completely at random. Due to the distributed nature of the data collection leading to our dataset, it was not possible to determine exactly what forces were behind data missingness and correct for them. We ultimately decided to adopt a complete-case analysis (deleting all missing values) approach, accepting the bias from this approach as a less disruptive alternative to imputing the missing values.

Our logistic regression, decision tree, and random forest models were all trained and tested using the caret package in R. We performed an 80/20 split on our dataset to build a training and test set so that we could validate our model performance on new data. Our training dataset consisted of 64,505 observations and the testing set used for evaluating our models’ performance contained 16,126 observations. In order to improve the predictive performance of our models, we used 10-fold cross validation with 3 repeats when training our models in caret. Additionally, since our diabetes dataset had an uneven class distribution, 38,385 readmitted and 42,246 not readmitted, we applied different sampling methods for balancing these class memberships. These methods included upsampling and down-sampling techniques available through caret as well as the use of the ROSE and SMOTE packages as synthetic and hybrid sampling approaches. These validation parameters, cross validation and class sampling, were defined using carets’ trainControl function as named list objects such as ctrldown. Each sampling method selected for dealing with our class imbalance was assigned to its own ctrl object. We trained different logistic regression, decision tree, and random forest models using our ctrl objects defined for each chosen method of dealing with class imbalances. The model that has the highest level of accuracy was selected as the final model for our analysis. The best performing method we selected to balance the class distribution of our final models was down-sampling. This sampling method was selected because it produced the model with the best predictive performance and used only actual observations when training the model.

As part of our model building we also attempted to build a predictive model using support vector machines. Our aim with using this modeling technique was to improve our predictive accuracy measure since support vector machines are more adept at handling messy data compared to our other modeling techniques. However, due to the size of our dataset we did not have the computing performance power necessary to build this model. Despite our inability to produce the support vector machine model, we have included the R code (shown below) that can be used to build this model when additional resources become available.

m.svm <- train(readmittedbinary ~ .,

trControl = ctrl,

metric = "ROC",

data = diabetesTrain,

method = "svmRadial")

p.svm <- predict(m.svm,diabetesTest)

confusionMatrix(p.svm,diabetesTest$readmittedbinary)

# Results and Performance

### **Results**

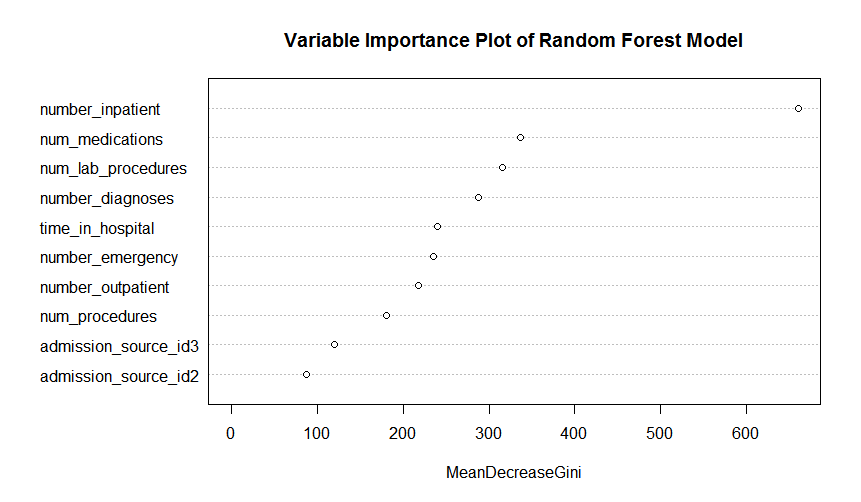
Based on the results of our chi-square analysis, all socio-demographic variables are significantly correlated to readmission risk. As shown in Figure 5, older patients have a higher likelihood of readmission. Only 43% of patients 30 or younger were readmitted, while 46% of patients between 30 and 70 and 50% of patients over 70 were readmitted.

*Figure 5.* Percent of readmitted patients within each age group.

As well as age both race and gender were significantly correlated to readmission risk. The Caucasian and African American populations has the highest proportion of readmissions as shown in Figure 6 below. Caucasian females had the highest proportion of readmissions at 49%.

*Figure 6*. Percent of readmitted patients by race.

All three of our models confirmed our second hypothesis that the number and type of prior clinical care visits is a predictor of readmission risk. Interestingly, the more inpatient, outpatient, and emergency visits a patient has significantly decreases the odds of readmission. We concluded that this result is likely due to that fact that doctors are aware that patients with a medical history of multiple prior clinical care visits need additional medical attention. Of the three types of clinical care visits, the number of inpatient encounter is the most significant to predicting readmission. Not only was inpatient encounter the most influential type of prior clinical care for predicting readmission, but it was the most significant predictor variable in all three of our models. According to our logistic regression model, each additional inpatient encounter decrease the odds of readmission by 66.27%. This significance is illustrated by Figure 7 which plots the variable importance of the top predictors of our random forest model in descending order.



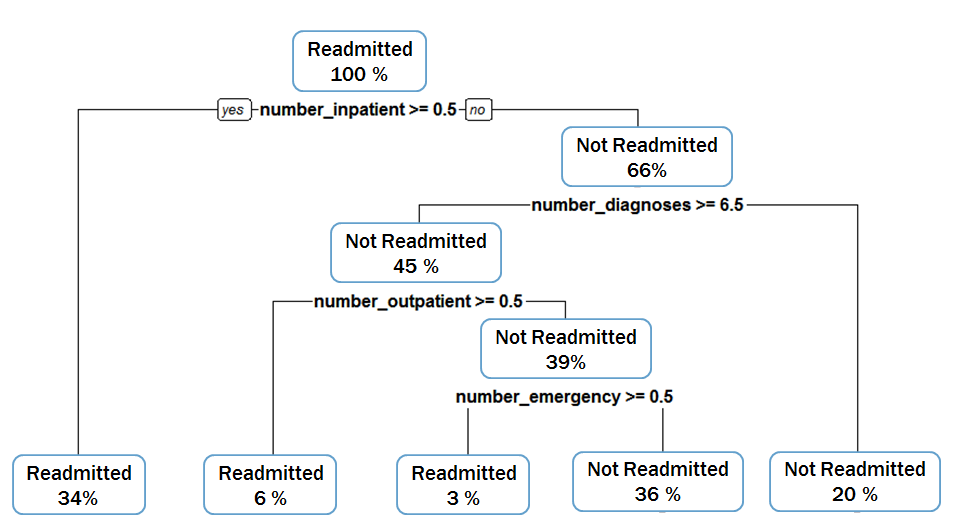
*Figure 7.* Variable importance plot of Random Forest mode.

Additionally, Figure 7 above indicates support for our third hypothesis that the number of lab procedures conducted during an inpatient encounter has a significant impact of readmission risk. Number of lab procedures was the third most important variable in our random forest model for predicting readmission. The results of our logistic regression model indicate that the odds of readmission decrease with each additional lab procedure that is conducted during an inpatient encounter.

Neither our logistic regression, decision tree, or random forest model indicated that insulin levels were a significant predictor of readmission. A single predictor logistic regression model was also used to look at the individual impact of insulin levels on readmission. While this model indicated a significant relationship between insulin levels and readmission, the model’s predictive performance on the testing data was extremely poor. The sensitivity was 0.2325 meaning our single-predictor model accurately predicted true patient readmission only 23.25% of the time. These results indicate that our fourth hypothesis that steady insulin levels significantly lower the likelihood of readmission is not supported. Similarly, changes in diabetes medications was not a key variable in any of our model and while single-factor logistic regression model indicated that changes in diabetes medications was significant, its predictive performance is poor. When used in logistic regression, changes in diabetes medications predicted readmission when a patient was actually readmitted only 49.20% of the time. Therefore, we concluded that there is not sufficient support to confirm our fifth hypothesis that changes in diabetes medications is significant to predicting readmission. Not only that, but the direction of impact we hypothesized was not reflected in the model. Our logistic regression model indicates that for patients with no change in diabetes medication during their inpatient encounter, the odds of readmission increase by 20.71%. Our hypothesis had predicted that a change in diabetes medications would increase readmission instead.

Our model building approach allowed for the identification of additional significant predictors of readmission as well. These included, number of medications, admission to hospital from other care centers, a primary diagnosis categorized as Other and number of diagnoses. Number of medications is shown as the second most significant variable for predicting readmission in our random forest model variable importance plot depicted in Figure 7 above. We also looked at variable importance for evaluating our logistic regression model where we identified two additional significant variables not represented in our hypotheses. The effects of these variables on readmission include the following; patients transferred from other care centers have a higher risk of readmission than patients admitted by referral and patients with a primary diagnosis categorized as other are more likely to be readmitted that patients whose primary diagnosis is categorized as circulatory. The final additional key predictor identified, number of diagnoses, was significant in all three of our models. It was the second most significant predictor of readmission in both our logistic regression and decision tree models. For each additional diagnosis recorded during an inpatient encounter the odds of readmission increases by 92.3%. Looking at Figure 8 below we can see the proportion of readmitted patients to not readmitted patients increases for patients with a greater number of diagnoses recorded during an inpatient encounter. Furthermore, Figure 9 shows that if a patient has greater than 6.5 diagnoses recorded during an inpatient encounter the patient is likely to be readmitted.

*Figure 8.* Number of diagnoses as a factor in readmission risk.

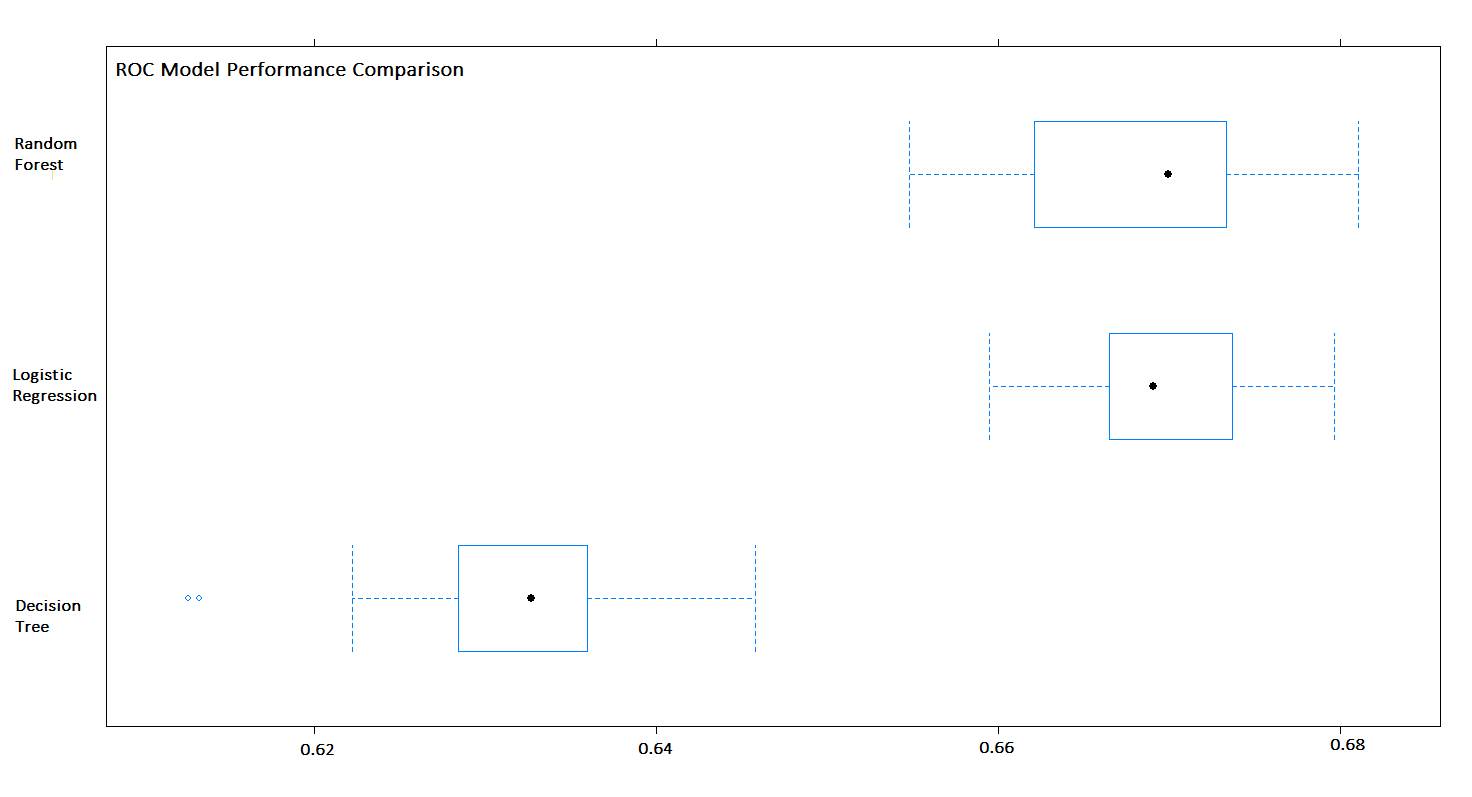
**

*Figure 9.* Final decision tree model.

**Model Accuracy Assessment**

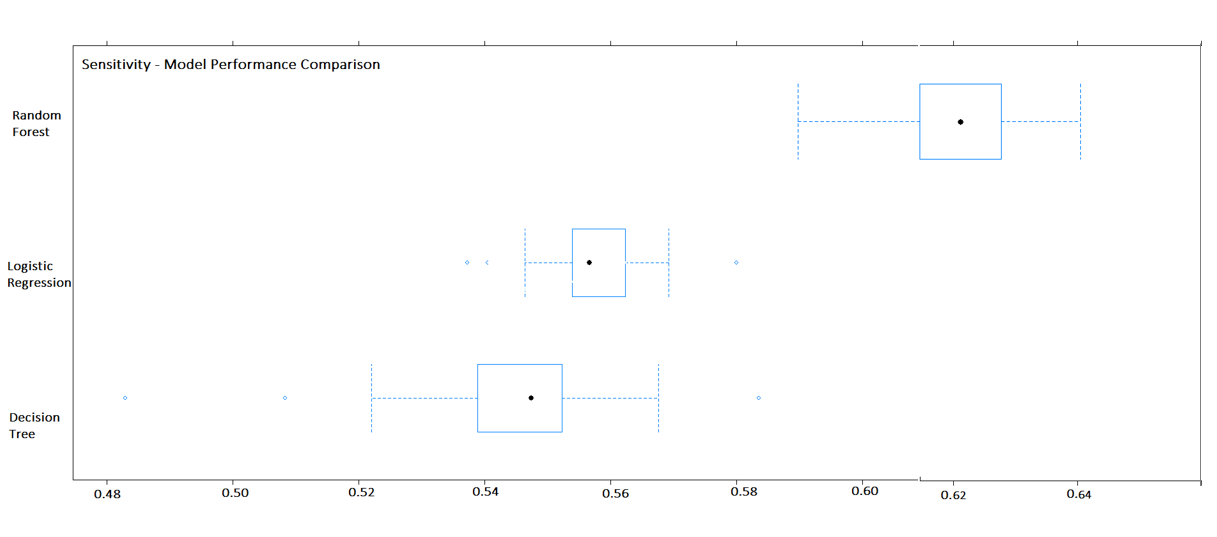
Our random forest model had the highest overall accuracy at 68.19% for predicting whether a patient was readmitted or not readmitted. Originally we’d found this accuracy level to be 63.58%, but during a final review of our code we found inconsistencies between the testing and training data being used in our random forest modeling. After resolving this issue with our code, the accuracy level of our random forest model improved to 68.19%. As seen in Figure 10 below, this accuracy level was followed by our logistic regression model at 62.65% and decision tree at 62.03% overall accuracy.

Our random forest model met our success criteria for overall accuracy greater that 65% in predicting readmission. While our other models did not quite meet our goal of 65% overall accuracy for predicting readmission, their identification of multiple significant predictor variables provides valuable information for identifying readmission trends. The similarities in the key predictor variables in each of our models enforces the significance of our findings and identifies areas of focus for process improvement. Currently, there are no standards in place for



*Figure 10.* ROC model comparison plot.

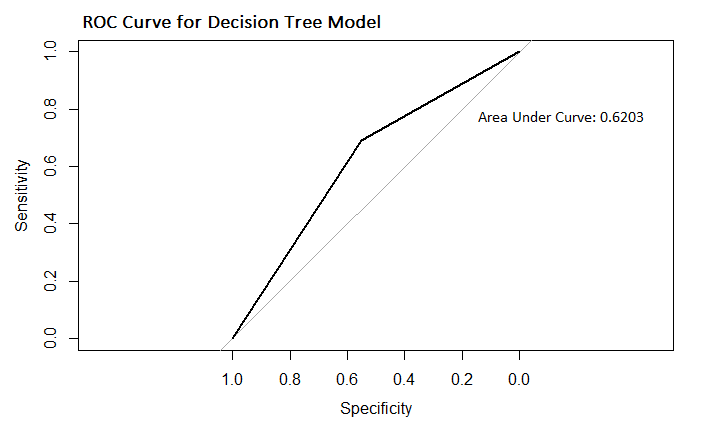
hospitals to measure likelihood of readmission of diabetes-related inpatient encounters. Since diabetes-related inpatient encounters account for a higher proportion of all inpatient visits, a model with 68.19% accuracy is better than random-chance predictions of readmission. Of our three models, random forest most accurately predicted readmission when the patient was actually readmitted, 64.54% of the time. This was followed by the logistic regression model which accurately predicted true readmission 56.52% of the time and then our decision tree at 55.20%. Figure 11 below further illustrates this sensitivity analysis comparison. Random forest had the highest overall accuracy and sensitivity, but its’ specificity was the lowest at 67.27%. Logistic Regression had highest specificity in that it accurately predicted non-readmittance when the patient actually was not readmitted 68.22%, followed by decision tree at 67.42%. Despite its’ slightly smaller specificity, our random forest model is still considered the best model since we are interested in predicting the actual class, readmission. These sensitivity measures and specificity measures use a cutoff value of 0.5. Since we are more interested in identifying patients that were readmitted we may want to consider lowering the cutoff value for



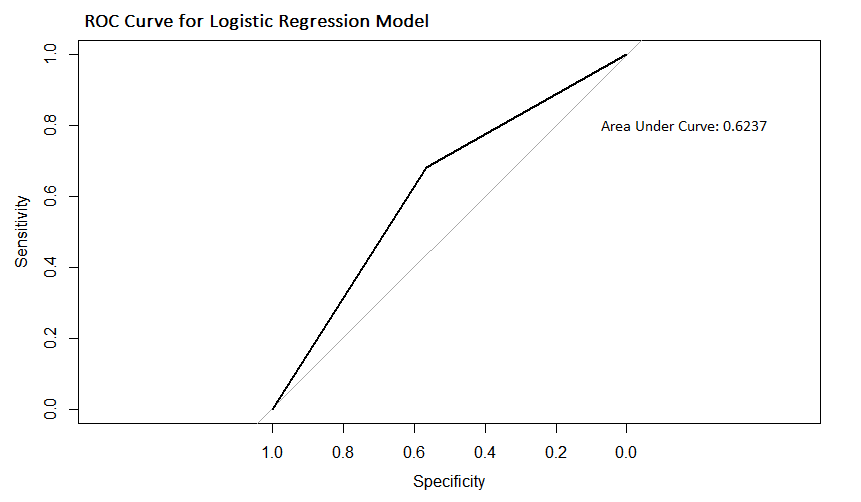
*Figure 11.* Sensitivity model performance comparison.

our model. This would mean that more inpatient encounters would be classified as readmitted, therefore our model would be better at capturing true readmissions.

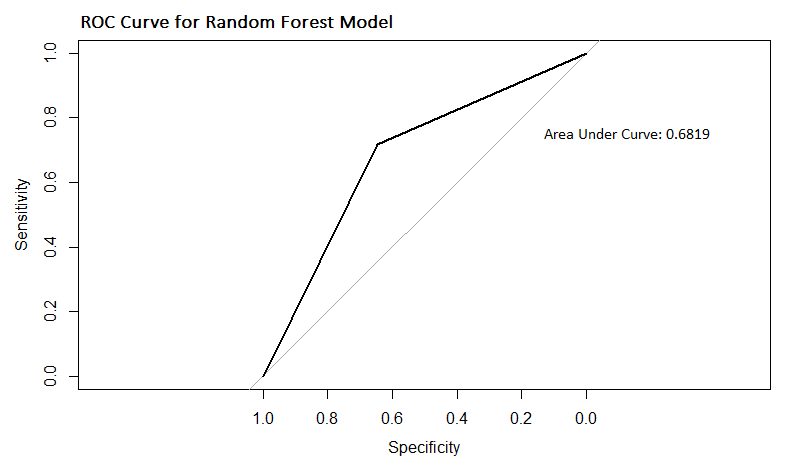
We built ROC curves to further evaluate the performance of our models in predicting readmission of diabetes-related inpatient encounters. Looking at Figures 12-14, we can see that all of our models outperformed random chance in predicting readmission of diabetes related inpatient encounter. While the lift is clearly evident in all three of the ROC curves, the random forest model outperformed the logistic regression and decision tree models. It has the highest area under the curve 0.6819, followed by logistic regression at 0.6234 and decision tree at 0.6203.



*Figure 12.* ROC curve for Decision Tree model.



*Figure 13.* ROC curve for Logistic Regression model.



*Figure 14.* ROC curve for Random Forest model.

# Discussion and Recommendations

Given the extremely high costs associated with hospital readmissions that were outlined in our original proposal, accurate predictions of readmission risk can make significant positive impacts on healthcare affordability and efficacy. The main overall recommendation that emerges from our research is to continue funding the development of readmission risk predictive models. Although we have already hit our success criteria by building a model with 68.19% accuracy and identified several factors that are significant predictors for risk of readmission, more development can still prove beneficial. Even slight increases to accuracy rate can translate into very large cost savings in the long run.

One aspect of this continued development should be realized in the form of a pilot program implemented in a small number of hospitals. We would ask hospital staff to run every diabetes patient’s information through our model which would then flag high risk patients for closer examination. At that point, a healthcare professional would attempt to provide preventative care aimed at avoiding readmission. By testing the model in real-world conditions we can identify and correct any issues that our initial investigation did not uncover. Additionally, testing on a small scale ensures that fixes can be deployed quickly and more cost effectively. This pilot should also provide insight into some of the reasons hospitals end up with missing data when inputting information about patient encounters into their system.

Our final recommendation is to provide funding for the development of improved integration between the electronic medical records (EMR) of healthcare organizations. One of the challenges we faced during our analysis was the issue of missing values. Due to the fractured nature of healthcare record keeping it was not possible for us to trace the origins of missing data to uncover whether systematic technological errors were responsible, systematic procedural errors were responsible, or if the values were missing completely at random. By developing some form of central data warehouse that is compatible with any healthcare information system we aim to improve the amount and quality of data available for all researchers. Standardized methods for capturing healthcare data should reduce instances of missing data and help improve model performance.

# References

Centers for Disease Control and Prevention. (2015, December). *Incidence and Age at Diagnosis*. Retrieved from https://www.cdc.gov/diabetes/statistics/incidence\_national.htm

Hines, A. L., Barrett, M. L., Jiang, H. J., & Steiner, C. A. (2014, April). *Conditions With the Largest Number of Adult Hospital Readmissions by Payer, 2011*. Retrieved from <http://www.hcup-us.ahrq.gov/reports/statbriefs/sb172-Conditions-Readmissions-Payer.pdf>