Draft Report of Final Project: Heritage Health Network Competition

Shivam

Model Developed: ANN

sjuneja@mtu.edu

Chathura

Model Developed: GBM

cjgunase@mtu.edu

Kyle

Model Developed: Reg.Trees

kajohnst@mtu.edu

**ABSTRACT**

This report is to describe the formulation and solution of the Heritage Health Price Kaggle competition. The work distribution between the group member and what each member’s contribution to the final outcome will be specified. Intuitively it would seem that by using the health history of a population it is possible to offer better-than-random predictions of the likelihood that a member of that population will need to go to the hospital in the following year.  Our project is to use the Heritage Health Prize data in order to predict the number of days the patient will spend in hospital next year.  The necessary steps include preprocessing the data, creating a data set suitable for the analysis, predicting the number of days spent in hospital during year two based on the claims of year-1 data using regression models. Finally we will build various regression models and compare the predicted values and actual values for accuracy and error rate. Comparison of the regression models will also be discussed in this report.

**Categories and Subject Descriptors**

Health Care Informatics, Electronic Medical health Records

**General Terms**

Electronic medical records, Regression, data analysis

**Keywords**

Keywords are your own designated keywords.

# INTRODUCTION

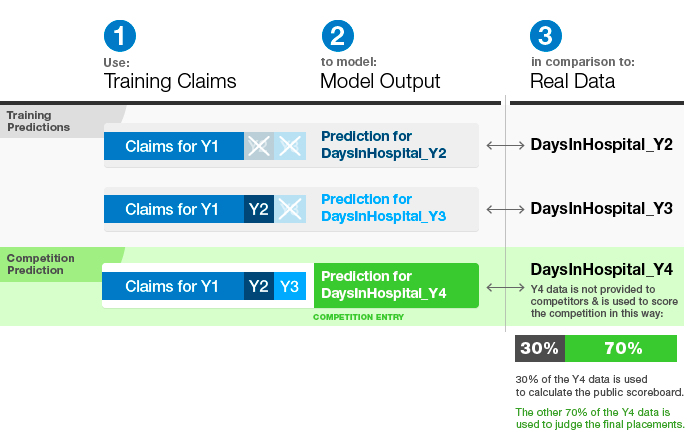
In the United States of America more than 100.7 million patients are admitted to hospital emergency rooms each year, and 136.3 million patients are admitted to outpatient clinics according to the CDC [[link](http://www.cdc.gov/nchs/fastats/hospital.htm)]. It is desirable to keep those patients healthy both for the humanitarian and cost reasons, with $1,500-$2,000 saved each day that the patients are kept out of the hospital [[link](http://www.beckershospitalreview.com/lists/average-cost-per-inpatient-day-across-50-states-in-2010.html)]. The Heritage Health Network (HHN) has provided an anonymized health history dataset on the Kaggle platform so that researchers can develop predictive models to indicate which patients are likely to spend time in the hospital, and how many days they will be admitted for. These predictions will allow clinicians to prioritize their resources to intervene in the most serious and costly cases, helping to alleviate strains on the healthcare system.

# BACKGROUND

Many methods have been researched in order to predict the length of a patients stay in a hospital [[link](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1067345/), [link](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6406418&tag=1)]. Commonly, the health records of a patient population has many missing values, and this must be accounted for by the researcher beforehand. With preprocessed data, researchers cluster the data set to create different training sets and a uniform test set. The most common predictive models used to predict length of stay are Random Forests, gradient boosting machines, support vector machines, and artificial neural networks. While these methods

**3.1 DATA DESCRIPTION AND PROBLEM UNDERSTANDING**

The data provided by the HHN is the same information that would be recorded in the patient's health record. The major sections of the record include the patient's general provider information including their primary caregiver, provider, and vendor. In addition, data is included on the logistical side of each instance of a claim including the year of claim, the delay in payment, the specialty visited, etc. Finally, data is recorded on the patient's medical history, including information on the number of drugs they take and of their lab tests. Detailed information on each variable can be found on the Kaggle page [[link](https://kaggle2.blob.core.windows.net/competitions-data/hhp/2496/Data_Dictionary_release3.pdf?sv=2012-02-12&se=2015-04-07T03%3A05%3A18Z&sr=b&sp=r&sig=F3aNuxxPTlYrYPGfnB5NRfONwKr3jZflIljS%2B0OPfoQ%3D)]. Intuitively it seems that using this information you would be able to make reasonably accurate predictions on which patients will have the most need of hospital stays in the coming year. For example, if a patient is diagnosed with a serious health concern, say heart disease, historical data should show that patients with that condition have regular visits to their specialty physician to manage their disease. In addition, it would be expected that the clinic and physician visited by the patient would have some bearing on their overall health, depending on the experience and expertise. Other weak patterns surrounding the relationship between hospital stays and the many variables provided should result in more accurate final predictions on length of stay.



## 3.2 PREPROCESSING AND DATABASE QUERY FORMULATION

The final output is number of days a patient spent in the hospital in the coming year given the claims data of this year. The original files are given in 6 different tables. Following is a brief description about each table and fields of each tables.

Members Table, which will include:

1. MemberID (a unique member ID)
2. AgeAtFirstClaim (member's age when first claim was made in the Data Set period)(CATAGORICAL)
3. Sex(CATAGORICAL)

This table comprise of unique Member ID which will be used in the result table to predict how many days each of this patient will spend in hospital next year.

The claims table is considered the main repository of medical claims and for each individual's claim is recorded when ever the patient receive a service from the hospital. There can be multiple entries from one patient.

Claims Table, which will include:

i. MemberID

ii. ProviderID (the ID of the doctor or specialist providing the service)

iii. Vendor (the company that issues the bill)

iv. PCP (member's primary care physician)

v. Year (the year of the claim, Y1, Y2)

vi. Specialty (CATAGORICAL)

vii. PlaceSvc (CATAGORICAL)

viii. PayDelay (NUMERICAL)

ix. LengthOfStay (CATOGORICAL)

x. DSFS (CATAGORICAL)

xi. PrimaryConditionGroup (CATAGORICAL)

xii. CharlsonIndex (ORDINAL)

xiii. ProcedureGroup(CATAGORICAL)

xiv. SupLOS

Labs Table, which will contain certain details of lab tests provided to members.

RX Table, which will contain certain details of prescriptions filled by members.

DaysInHospital Tables - Y2 and Y3, which will contain the number of days of hospitalization for each eligible member during Y2 and Y3 and will include:

i. MemberID;

iii. DaysInHospital (the number of days in hospital Y2 or Y3, as applicable).

These tables were aggregated to create a final dataset which will be used to build predictive models. Dummy variables were created for categorical variables. This has expanded the dataset to 147 predictors.

## 3.3. DATA EXPLORATION

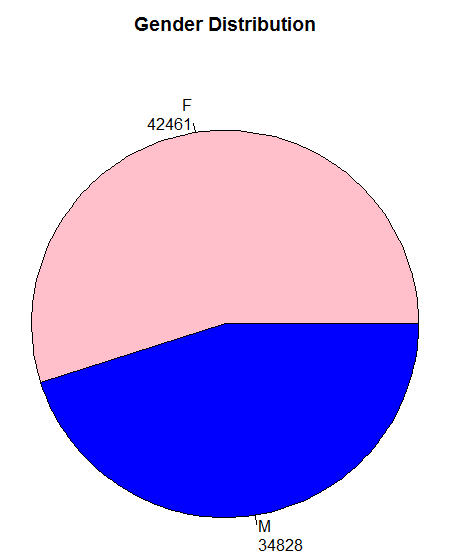
Exploratory Data Analysis:

We study different variables pairwise:

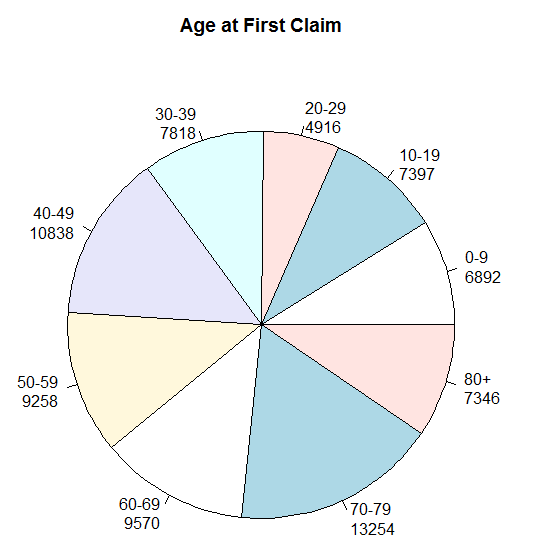
There are 77289 Unique members and majority of these members are females

The age at first claim has 9 categories i.e 0-9 years, 10-19 years etc.

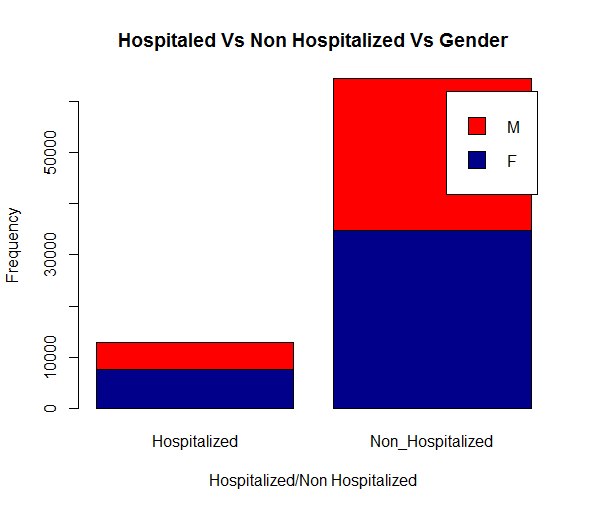
Majority of the members are non hospitalized



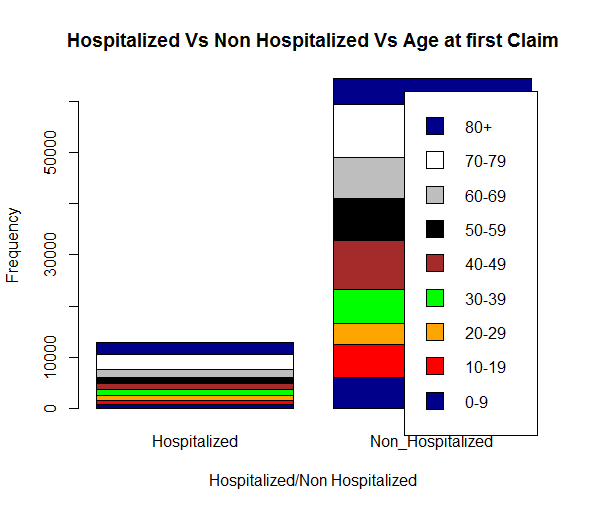
This pie chart shows the distribution of Gender. It is clear that there are more female patients than male patients



This pie chart depicts the age of the patients at first claim. Age group of 70-79 years has the maximum number of claims



The bar-plot displays the categorical classification of male and female patients/members who were hospitalized. We can clearly see that majority of the hospitalized members/patients are females



# Predictive Model Building

This phase workload will be distributed between each group member to train and test one or more models. Each member will produce a predictive model Regression Models, explain in detail how they work and produce graphs and evaluation plots.

## 4.1 Model Evaluation Criteria

The response variable of this model is “Number of Days the Patient will spend in the hospital next year”. Because this is a discreet numerical variable following transformation was done to convert the days in to log+1 scale.

Where i is a patient.

After the prediction the predicted values will be converted to original scale using the following transformation.

When building regression models modified Root mean squared error will be used as the minimization function.

A Best individual model will be selected by the minimum RMSE.

## Chathura - Gradient Boosting Machines

## Shivam - Neural Networks, MARS

## Kyle – Bagged Trees

# ENSEMBLE

After each group member completes Compare each model and build an ensemble predictor compare again.

The bar-plot displays the categorical classification of members/patients who were hospitalized according the age group. We can clearly see that majority of the hospitalized members/patients belong to age group of 70-79 years.