Predicting Time and Cost of Hospital Readmission

Stacey Newman Diksha Garg Zeinab Mahdavifar
Center for Data Science
University of Washington Tacoma
Advisors: Prof. Martine De Cock, Dr. Shanu Sushmita

I. Introduction

Medical care is the basic requirement. It is very important that every person gets a good care at optimum price. Due to many influential factors and medicare plans, this cost has significantly increased over the years. This cost needs to be managed in a way to be accessible to every individual. Among these costs, some are not only manageable, but preventable. Hospital re-admissions is one of these common and expensive concerns.

According to recent studies, there are about 18% of medical beneficiaries who get re-admitted within 30 days, increasing the readmission cost up to 17 million annually in the United states[6]. Re-admission is a problem which contributes to additional costs for the health care system as well as added risks for patients. Patients diagnosed with Congestive Heart Failure (CHF) are highly prone to readmission within 30 days. It's been surveyed that CHF alone covers nearly 25% of the re-admissions. As there are financial penalties for hospitals associated with 30-days readmission, preventing these penalties in an effective and efficient manner is of significant concern to hospital administrators and healthcare providers. When it comes to the cost of health care, according to World Health Organization (WHO), US has the highest cost of health care compared with other developed countries. To bring this cost down to a manageable figure, various healthcare reform policies are proposed. As the researches [11] indicates, re-admissions are expensive but preventable. So, if the risk of readmission is reduced, significant decrease in healthcare cost can be achieved.

Predicting risk of readmission is a challenging process that involves clinical features, demographic and social factors, disease and health parameters, the quality of care provided by hospitals, and large number of other factors specific to patients and the health care provider. The important features that contribute toward this risk prediction must be mined from this vast pool of data to build accurate machine learning models that can predict a patient's risk of re-admission. Using all these features, accurate models are proposed to predict risk as well as the most important features that contributes towards the calculation. The problem becomes even more complex when the data requires significant pre-processing and normalization. Predicting the cost of these re-admissions adds another layer of complexity. The distribution of medical costs in [2] displays a long right tail (due to extreme bias towards a specific class), and makes attempts to build accurate predictive models particularly challenging.

The goal of this project is to propose a dual predictive tool that utilizes healthcare data to predict time and cost of hospital re-admission. The dual predictive model will help in analysing the risk associated with re-admission of the patient i.e. the number of days in which he might be re-admitted to the hospital, along with the associated cost that would be incurred due to readmission. This is a supervised, multilabel, multi-class classification problem. For analysis and effective prediction, the cost has been divided into three categories- high,medium and low. If the cost is less than $14,789, it is assigned under low cost category, cost values between $14,789 and $29093 are medium and cost values greater than $29,093 are assigned to high. In the same way, for risk prediction three categories are defined based on days of re-admission. Categories are 30, 60 and 90 days re-admission. Machine learning algorithms- Decision trees, Naïve Bayes and Support Vector Machine, are used for classification and prediction models. Washington state Inpatient data is used for this experiment.

The preceding sections are organized as follows: Section 2 introduces with related work available in this field. Sections 3 gives a brief overview of dataset followed by Section 4 which presents the feature selection results. Section 5 describes the problem statement in detail. Methods and algorithms used for analysis are covered in section 6. Section 7 explains the evaluation measures used. Results are presented in section 8 followed by conclusion and future work in section 9 and 10 respectively. All the tables and charts generated during the experiment are placed in three appendices at the end.

II. Related Work

The previous work done in this area has addressed the problems of risk of readmission and cost prediction separately. In [7] K. Zolfaghar et. al. proposed a big data solution for predicting 30 days risk-of-readmission for congestive heart failure patients. The study results leverages big data infrastructure for information extraction, information integration and predictive modelling. In [2] D. Bertsimas et.al. present cost prediction of health care expenses using health insurance claims data of about 800,000 beneficiaries over 3 years. The results of this work conclude that medical information of high cost members helps achieve more accurate predictions, thus motivating the integration of many data sources.

In this work, we investigate different approaches and machine learning models to predict both the readmission time window of a patient as well as the cost that will be incurred due to this readmission.

III. Dataset and Features

The data used for the experiments is ”State Inpatient Databases (SID)”[3] for the State of Washington. The SID database is provided by the Healthcare Cost and Utilization Project (HCUP), which is a family of health care databases and related software tools and products developed through a Federal-State-Industry partnership. It is sponsored by the Agency for Healthcare Research and Quality (AHRQ). The original dataset consists of inpatient discharge records from community hospitals in the State of Washington with all-payer, encounter-level information from 2009 to 2012. The data consists of 650,000 records per year corresponding to 480,000 beneficiaries with over 1000 attributes. We are using a subset of the SID dataset with preprocessed data specifically about congestive heart failure (CHF) patients to predict the time and cost of next readmission. The initial data was formatted into 834 columns. The cohort of patients admitted for CHF that we extracted consisted of 4,500 rows. The features used are briefly summarized in Table I.

Feature Number	**Description**
1-3	Demographic Information
4-7	Historical/Admission Information
8-269	Diagnoses
270-500	Procedures
501-529	Comorbidities
530-804	Revenue Codes
805-834	Utilization Codes
Total Number of Instances: 4,500	

Table I

Summary of the full feature set in the SID data.

A. Initial Feature Space

We also trained models using a specific subset of the feature set. The 8 features chosen composed of basic information a patient may know upon discharge and excluding complex medical information such as revenue codes. We also used a slightly smaller cohort of CHF patient admissions (3,672 admissions). A brief summary of the dataset used in the initial feature selection and cohort extraction is summarized in Table II.

Feature	**Description**
1	Primary Diagnosis
2	Diagnosis Related Group
3	Primary Procedure
4-6	Demographics
6-8	Historical/Admission Information
Total Number of Instances: 3,672	

Table II

Summary of the feature set representative of information known by individuals.

Previous work presented in [8] provides individuals, not just healthcare entities, a way to access these types of predictive models. This could help individuals take charge of their own finances and personal health.

B. Feature Selection

Feature selection is a solution to one of the most fundamental challenges in machine learning: the curse of dimensionality. Although high speed computation enabled us to deal with high-dimension data much easier, we still need feature selection to boost the accuracy of the output. One main reason for feature selection apart from reducing the dimensionality is the fact that the increasing number of features, increases the probability of missing attributes. Obviously we can not delete all columns with missing attributes, as it ends up with shortage in samples. So, it is wise to lose features which do not impact prediction rather than samples with more valuable information. Another potential benefit of feature selection is possibly decreasing over-fitting by the models. This allows models to be more general and potentially more flexible for predicting new data instances that deviate from the instances seen in training. This benefit is most noticeable in models similar to decision trees. We define feature selection formally as follows:

Let function f demonstrate the true relationship between the input set $X = \{x_1, x_2, ...x_M\}$ and output Y. It is common that output Y is not determined by the complete set for features, but only a subset of X like $\{x_1, x_2, ..., x_m\}$ where $m < M$ contains the features determining function f. The irrelevant features induce a great deal of unnecessary computational cost and potentially decrease the accuracy of models produced by machine learning algorithms such as Naïve Bayes.

1) Approaches: There are a plethora of methods and approaches for performing feature selection. The brute-force version that comes to mind is exhaustively evaluating all possible features and their combinations to find the best subset. This method comes at a great computation cost, however. We performed feature selection methods through a greedy approach. These methods can be classified into two main groups: filter-based and wrapper-based approaches. In filter-base approach, some correlation analysis is used on each feature separately. On the other hand, wrapper-base works within a machine learning algorithm and at each iteration, base on forward or backward wrapping, adds or removes a feature to improves the model performance. Since each of these methods come with their own pros and cons, both approaches were used to identify best features.

Filter-based feature selection provides an understanding of how the attributes affect the predicted value using some similarity measure. Not having the domain knowledge about the data is common in machine learning projects, and filter-based feature selection provides knowledge about the actual relationship between variables. Still, using filter-based approach alone considers each attribute individually, resulting in unstable or inaccurate prediction[4]. In order to identify the best features examining each individually is not enough, because it will not capture how attributes affect the prediction as a set. One probable scenario that is not captured by filter-base approach is multi-collinearity issues, that happens when the predictions correlates with two or more attributes that are also highly correlated (one can be predicted from the others). Using wrapper-based approach, we can handle this issue automatically, but since it will not give any further knowledge about variables, we need to incorporate both approaches in parallel.

2) Filter-based Feature Selection: For using filter-based, we ran experiments in AzureML[9]. We used two separate metrics to determine the importance of a feature: Mutual Information (MI) and Chi Squared(X^2). Mutual information is a measure of two independent variables' mutual dependence. Here we chose MI to characterize both the relevance and redundancy of variables, and it is calculated according to the following formula:

$$MI(X;Y) = \sum\sum p(x,y)log\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

where $p(x,y)$ is the joint probability distribution function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y respectively. We used MI to find the degree of relevance of a feature to the readmission time window, the readmission cost, and both time and cost. Figures 1, 2 and 3 in Appendix B shows the attributes sorted by their MI score.

We also used the chi squared (CHI) measure(X^2) for filter-based feature selection. This statistical measure was used to evaluate how likely it is that any observed difference between variables arose by chance. Formula(2) was used to determine CHI measure, where E_i stands for the expected value and O_i is the observed value of attribute i in the experiment. The results from applying this measure on attributes are shown in figures 4, 5 and 6 in Appendix B.

$$X^2 = \sum_{i=1}^{n}\frac{(O_i - E_i)^2}{E_i} \tag{2}$$

The final result of filter-base approach is summarized in tableX, XI, and XII in Appendix C.

3) Wrapper-based Feature Selection: For wrapper-based feature selection, we used the "FSelector" package in R[10]. Using this package we did feature selection using three metrics: information gain, gain ratio, and symmetrical uncertainty. They are entropy based algorithms which find weights of discrete attributes based on their correlation with continuous class attribute. These algorithms yielded the best results needed for this experiment.

These metrics are defined in formulas (3),(4), and (5) where $H_{(x)}$ stands for entropy.

$$Information\ Gain = H_{class} + H_{attribute}H_{class|attribute} \tag{3}$$

$$Gain\ Ratio = \frac{(H_{class} + H_{attribute}) - H_{class|attribute}}{H_{attribute}} \tag{4}$$

$$Symmetrical\ Uncertainty = \frac{2(H_{class} + H_{attribute}) - H_{class|attribute}}{H_{class} + H_{attribute}} \tag{5}$$

For each measure, we found the 20 most relevant features for cost, readmission time, and combination of both time and cost. The results from using "FSelector" R package is shown in tables XIII, XIV and XV in Appendix C.

4) Final Feature Selection: To use the results from filter-base and wrapper-base approaches, we used the results from each approach and performed a majority voting on all results, such that we only chose the attributes that were at least chosen by 3 metrics out of 5(Chi Square, Mutual Information, Information Gain, Gain ratio, and Asymmetric Uncertainty). Then the top 20 were chosen as the final feature set to use for Decision Tree, SVM and Naïve Bayes. Also, the same feature set was used to evaluate the best model. These 20 features are shown in table III and include attributes on historical and admission information of patients, their age, diagnosis, and procedures, as well as their revenue code.

Final 20 features to test the final model
Total charges, Cumulative Cost, Age, Length of Stay, Revenue921, Revenue361, Revenue272, Revenue391, Revenue440, Diagnosis249, Ultrasound, Revenue402, Diagnosis2, Revenue302, Diagnosis157, Procedure50, Diagnosis109, Procedure216, Procedure223, Procedure54

Table III
FINAL FEATURES: SELECTED BY MAJORITY VOTING BETWEEN RESULTS OF FILTER-BASE AND WRAPPER-BASE. ONLY FEATURES WITH 3 VOTES AND HIGHER ARE SELECTED.

Table IV provides an overview of the dataset used after feature selection.

Feature	**Description**
1-3	Historical/Admission Information
4	Age
5-9	Diagnosis Groupers
10-13	Procedures
14-20	Revenue Codes
Total Number of Instances: 4,500	

Table IV
SUMMARY OF DATASET AFTER FEATURE SELECTION.

IV. PROBLEM DESCRIPTION

The goal of our work is to, at the time of a patient's discharge, be able to predict the time window in which the same patient will be readmitted and what the cost of that readmission will be. In our work we only consider the cohort patients who were first admitted for congestive heart failure. We approach this problem as a supervised, multilabel, multi-class classification problem. We observe N training instances composed of two vectors, (X_i, Y_i) for $i = 1, 2, ..., N$, to learn a model that can map an unseen X_i to the appropriate Y_i. The vector $X_i = \{x_{1_i}, x_{2_i}, ..., x_{m_i}\}$ consists of m demographic, claims, and clinical data points (features) available after discharge of a patient i. We use this information to predict vector $Y_i = \{y_{1_i}, y_{2_i}\}$, which consists of two labels - one predicting the time window of the next readmission, y_1, and a second predicting the severity of that admission's cost, y_2.

We consider three classes for the time of readmission label: 30 days, 60 days, and 90 days. A label of 30 days means that our model predicts the patient will be readmitted within 30 days of discharge. Similarly, 60 days would indicate a readmission within 60 days but after 30 days of discharge and 90 days would indicate a readmission within 90 days but after 60 days of discharge. No instances of readmission after 90 days are included in either training or testing data and are therefore not considered.

We also consider three classes for the cost of readmission label: high cost, medium cost, and low cost. A label of high cost would indicate the patient's readmission cost is greater than \$29,093, a label of medium cost would indicate a cost less than \$29,093 but greater than \$14,789, and a label of low cost would indicate a cost less than \$14,789.

To illustrate the meaning of these labels, consider a patient p who is admitted to the hospital for congestive heart failure. Patient p is then discharged on January 2 but is readmitted again on February 10 of the same year. This readmission costs a total of \$20,000. Given the demographic, claims, and clinical data available on January 2, we would like our model to output $Y_p = \{60\ days, medium\ cost\}$. Table V gives the distribution Y present in our full dataset:

Label	Percent of Data
90 days, high cost	8.89%
60 days, high cost	6.73%
30 days, high cost	26.33%
90 days, medium cost	8.22%
60 days, medium cost	6.18%
30 days, medium cost	17.53%
90 days, low cost	6.18%
60 days, low cost	5.07%
30 days, low cost	14.87%

Table V
LABEL DISTRIBUTION IN THE FULL DATASET.

Table VI shows the distribution of Y present in a specific portion of the CHF cohort extracted after our first round of feature selection.

Label	Percent of Data
90 days, high cost	3.27%
60 days, high cost	7.11%
30 days, high cost	22.88%
90 days, medium cost	3.59%
60 days, medium cost	7.46%
30 days, medium cost	22.49%
90 days, low cost	3.49%
60 days, low cost	7.65%
30 days, low cost	22.09%

Table VI
LABEL DISTRIBUTION IN THE EXTRACTED SUBSET OF THE FULL DATASET.

V. METHODS

This section will give an overview of the approaches, models, and baselines used in this research.

A. Baseline Algorithm

We include two baselines in our results: the majority baseline and the random baseline. We seek to learn about the performance of different models and approaches via comparison to these baselines. If our model is able to out-perform the random baseline, then we know it has performed better than random guessing. If our model is able to out-perform the majority baseline then we know it has performed better than always assuming the most common Y value.

1) Majority Class: The majority class baseline is calculated by always predicting the Y that is seen most often in the training data. For example, if the dataset consists of 4,000 training instances, the majority class model selects the Y seen most often in those 4,000 instances, say Y_{mc}. Then, when the model is given 500 unseen test instances, it predicts Y_{mc} for each of these instances.

2) Random Class: The random class baseline is calculated by randomly choosing the prediction value of Y from the set of all Y seen in the train data. For example, if the dataset consists of 4,000 training instances, the random class model creates a set S of Y values seen in these 4,000 instances. No Y value is present more than once in S. Then, when the model is given 500 unseen test instances, it randomly samples, with replacement, 500 values from S. The model then predicts these 500 samples as the corresponding Ys for the unseen instances.

B. Binary Relevance Method

A simple approach to multi-label classification is the Binary Relevance (BR) method. In this approach, separate models are trained for each label independently. The predictions from the separate models are then aggregated to produce the final prediction. In our case we train two multi-class classifiers - one to predict the time window of the next readmission and a second to predict the cost of that readmission. The first model takes the full vector X_i for a patient i as input and produces readmission time prediction y_{1_i}. The second model will similarly take X_i as input but will produce the cost prediction y_{2_i}. These two predictions will then be aggregated to produce the final prediction $Y_i = \{y_{1_i}, y_{2_i}\}$.

We investigate three popularly used classification algorithms with the binary relevance approach - Support Vector Machines (SVM), Naïve Bayes classifiers, and Decision Trees. When using the decision tree base classifier we also tune the complexity parameter (cp) to three different values - 0.01, 0.001, and 0.0005. A higher complexity parameter results in fewer splits to the decision tree. More information on the complexity parameter can be found in [1]. It can be noted that these models in the Binary Relevance approach will not consider any label dependency that may exist between y_1 and y_2.

C. Label Powerset Transformation (LP)

The label Powerset (LP) transformation approach transforms the multi-label problem into a single-label problem. This approach, unlike binary relevance, considers label dependencies. Label Powerset transformation is performed within our problem by combining our two three-class labels into one nine-class label. This is done by concatenating y_1 and y_2 for every possible Y to create our nine classes. Using the label Powerset approach we investigate the same set of base classifiers - Support Vector Machines, Naïve Bayes classifiers, and Decision Trees (with complexity parameter tuned to 0.01, 0.001, and 0.0005).

D. The Chained Classifier Model (CC)

The Chained Classifier model approach requires a chain of classifiers $C_1, C_2, ..., C_{|L|}$, where $|L|$ is the number of labels to predict. Each C_k in the chain is a classifier built to predict label y_k. Each C_k takes as input the full vector X_i for a patient i as well as predictions $y_{1_i}, y_{2_i}, ..., y_{k-1_i}$ output from the preceding classifiers $C_1, C_2, ..., C_{k-1}$. C_k then outputs a prediction y_{k_i} for the patient. In this case, $|L| = 2$, C_1 is a model that takes X_i for a patient i as input and predicts the time window of the next readmission for patient i (y_{1_i}), and C_2 is a model that takes X_i and the prediction y_{1_i} output from C_1 as input to predict the cost of patient i's next readmission (y_{2_i}). These predictions are then aggregated for the overall prediction $Y_i = \{y_{1_i}, y_{2_i}\}$. Using the chained classifier approach we use the same set of base classifiers - Support Vector Machines, Naïve Bayes classifiers, and Decision Trees (with complexity parameter tuned to 0.01, 0.001, and 0.0005).

VI. Evaluation Measures

In classification problems, the most popular metric for evaluating a model's performance is the accuracy measure. In the multi-label scenario, however, the equations used to calculate this metric must be refined to indicate the case where a model may be partially accurate. In our case, we would like to reflect instances where the model may have predicted one of the two labels correctly in our accuracy measure. To reflect these instances we use a metric proposed by [5] referred to as average accuracy. Average accuracy intends to evaluate how well the model correctly identifies all of the labels and is calculated as follows:

$$(6) \qquad Average\ Accuracy = \frac{1}{N}\sum_{i=1}^{N}\frac{L_p(i) \cap L_t(i)}{L_p(i) \cup L_t(i)}$$

Where N is the number of predictions made, $L_p(i)$ are the predicted labels for instance i, and $L_t(i)$ are the true labels for instance i.

We also include average precision, average recall, and F-1 measure as evaluation metrics. Average precision represents the percentage of instances predicted as a given Y value actually have that same Y value averaged across all potential Y values. Average precision is calculated as follows:

$$(7) \qquad Average\ Precision = \sum_{i=1}^{C}\frac{ActualY_i}{PredictY_i}$$

where $ActualY_i$ is the number of instances predicted as Y_i and also have a true value of Y_i, $PredictY_i$ is the number of total instances predicted as Y_i, and C is the number of potential Y values.

Average recall represents the percentage of instances that truly have a certain value of Y that were accurately predicted same as Y averaged across all potential Y values. Average recall is calculated as follows:

$$(8) \qquad Average\ Recall = \sum_{i=1}^{C}\frac{TrueY_i}{ActualY_i}$$

where $TrueY_i$ is the number of instances that have a true value of Y_i and were predicted correctly, $ActualY_i$ is the total number of instances that have a true value of Y_i, and C is the number of potential Y values.

F1 measure is then calculated as follows:

$$(9) \quad F-1\ Measure = \frac{2*(Average\ Precision)*(Average\ Recall)}{(Average\ Precision)+(Average\ Recall)}$$

VII. Results

From the results presented in Table VII, we can see that the majority baseline is extremely hard to improve upon. The larger we grew the Decision Trees, by decreasing the complexity parameter, the more our average accuracy decreased. In some scenarios, such as Label Powerset on the full feature set, the majority baseline was the best performer. Any learning algorithm that did not align with the majority baseline would overfit the training data, thereby decreasing the accuracy on the test data.

Naïve Bayes came up as a consistent poor performer, whereas SVM and smaller Decision Trees appear to challenge the majority baseline in most scenarios.

None of the learning models were able to achieve average accuracy more than 50%. In the appendix, Table IX displays the confusion matrix for the best performing model that was trained over the dataset of 4,500 instances. This model was trained using the Binary Relevance technique with SVM base models using the 20 features found during feature selection. In our experiments, this model results an accuracy of 42.9%. This matrix clearly shows that the model shows a distinct bias toward the majority time window of 30 days. While this model did improve slightly over the results from the majority baseline, there was no significant improvement since the majority risk label was routinely chosen.

In addition, we used same model on a completely unseen test data. The results are shown in Table VIII. There is a small improvement on the majority baseline for the test data, but these results are similar to those in Table VII.

Approach - Algorithm	**Average Accuracy**	**Average Precision**	**Average Recall**	**F-1 Measure**
Full Feature Set				
Majority Baseline	42.34%	2.93%	11.11%	0.0463
Random Baseline	25.33%	10.73%	11.08%	0.1090
Binary Relevance				
Naive Bayes	31.58%	14.30%	14.62%	0.1446
SVM	42.34%	2.93%	11.11%	0.0463
Decision Tree(0.01*)	42.59%	7.55%	13.19%	0.0961
Decision Tree(0.001)	37.93%	13.32%	12.90%	0.1311
Decision Tree(0.0005*)	35.00%	12.62%	12.53%	0.1258
Label Powerset				
Naive Bayes	31.13%	12.68%	13.21%	0.1294
SVM	42.34%	2.93%	11.11%	0.0463
Decision Tree(0.01*)	42.34%	2.93%	11.11%	0.0463
Decision Tree(0.001*)	40.78%	12.81%	12.92%	12.86
Decision Tree(0.0005*)	36.84%	12.16%	12.48%	0.1232
Chained Classifier				
Naive Bayes	31.64%	14.51%	14.66%	0.1458
SVM	42.34%	2.93%	11.11%	0.0463
Decision Tree(0.01*)	42.59%	7.55%	13.19%	0.0961
Decision Tree(0.001*)	37.72%	13.13%	12.77%	0.1294
Decision Tree(0.0005*)	35.01%	12.56%	12.51%	0.1253
Initial Feature Selection and Cohort				
Majority Baseline	41.02%	5.01%	10.94%	0.0687
Random Baseline	25.41%	10.43%	10.29%	0.1035
Binary Relevance				
Naive Bayes	43.25%	12.41%	12.76%	0.1258
SVM	46.45%	9.47%	14.31%	0.1140
Decision Tree(0.01*)	47.11%	9.55%	14.71%	0.1158
Decision Tree(0.001*)	44.70%	14.49%	14.20%	0.1434
Decision Tree(0.0005*)	41.43%	13.93%	13.69%	0.1381
Label Powerset				
Naive Bayes	42.37%	11.25%	12.46%	0.1182
SVM	46.06%	9.21%	14.03%	0.1112
Decision Tree(0.01*)	47.34%	6.81%	14.94%	0.0936
Decision Tree(0.001*)	46.30%	14.37%	14.85%	0.1461
Decision Tree(0.0005*)	43.64%	12.41%	13.79%	0.1307
Chained Classifier				
Naive Bayes	43.27%	12.31%	12.76%	0.1253
SVM	46.54%	9.49%	14.40%	0.1144
Decision Tree(0.01*)	47.11%	9.55%	14.71%	0.1158
Decision Tree(0.001*)	44.83%	13.35%	14.22%	0.1377
Decision Tree(0.0005*)	41.39%	13.87%	13.64%	0.1375
Final Feature Selection				
Majority Baseline	42.34%	2.93%	11.11%	0.0463
Random Baseline	26.2%	11.66%	11.76%	0.1171
Binary Relevance				
Naive Bayes	31.80%	11.49%	13.95%	0.1260
SVM	42.90%	7.54%	11.84%	0.0921
Decision Tree(0.01*)	42.66%	7.66%	13.21%	0.0969
Decision Tree(0.001*)	41.50%	12.59%	13.38%	0.1348
Decision Tree(0.0005*)	36.90%	12.83%	12.85%	0.1284
Label Powerset				
Naive Bayes	31.14%	10.50%	13.47%	0.1180
SVM	42.34%	2.93%	11.11%	0.0463
Decision Tree(0.01*)	42.34%	2.93%	11.11%	0.0463
Decision Tree(0.001*)	41.66%	12.22%	12.90%	0.1255
Decision Tree(0.0005*)	38.79%	12.76%	12.36%	0.1305
Chained Classifier				
Naive Bayes	31.92%	11.54%	13.89%	0.1260
SVM	42.73%	6.92%	11.54%	0.0865
Decision Tree(0.01*)	42.63%	7.66%	13.21%	0.0969
Decision Tree(0.001*)	41.27%	13.59%	13.25%	0.1342
Decision Tree(0.0005*)	36.96%	12.81%	12.76%	0.1278

Table VII

ALL RESULTS OBTAINED USING THE THREE DATASETS PREVIOUSLY OUTLINED AS WELL AS THE THREE METHODS OUTLINED AND THREE DIFFERENT BASE MODELS. *THE DECISION TREES WERE GROWN WITH DIFFERENT COMPLEXITY PARAMETERS NOTED IN THE TABLE.

Approach - Algorithm	**Average Accuracy**	**Average Precision**	**Average Recall**	**F-1 Measure**
Unseen Test Data				
Majority Baseline	42.07%	4.25%	5.83%	0.0602
Random Baseline	26.27%	11.03%	5.88%	0.1127
Binary Relevance				
SVM	43.73%	5.61%	11.67%	0.0750

Table VIII
RESULTS FROM THE BINARY RELEVANCE SVM MODEL EVALUATED ON THE TEST DATASET.

VIII. CONCLUSION

In this study, we examined 3 different machine learning algorithms to predict the cost and time of readmission for Cognitive Heart Failure patients in Washington State. We used SID dataset from HCUP containing 4,500 rows of patients' hospital discharge information with 834 attributes for each record. The algorithms used were Decision Trees (with 0.01, 0.001, and 0.0001 complexity parameters), Support Vector Machines(SVM), and Naïve Bayes. The study included experiments on the full feature data set (all 834 features and 4,500 rows), experiments with features after feature selection (20 features and all 4,500 rows), and experiments on a smaller dataset representative of individual information (8 features and 3,672 rows) using Binary Relevance, Label Powerset, and Chained Classification models. The best results before feature selection is with Decision Trees (0.01 complexity parameter) using the Binary Relevance model with 42.59% average accuracy. This was improved to 42.90% after feature selection, using Binary Relevance with SVM. Finally by running our best model (SVM) on the test data, we improved to 43.73% average accuracy. Our work has shown that the majority baseline is difficult to out perform with such a strong skew toward 30 day readmission.

The problem of predicting when a patient will be readmitted and the cost of that readmission remains an important problem and the difficulties outlined in this work reinforce that it is also an interesting one.

IX. FUTURE WORK

Considering the challenge of beating the majority baselines, there are plenty of opportunities for future work in this area. One approach may be to address to sparse vectors seen in the data. Many columns, especially the diagnosis, revenue codes, and procedures, were mostly 0. Investigating techniques used in topic modeling, such as singular value decomposition, and whether they could increase accuracy by decreasing dimensionality.

Another potential improvement which may increase accuracy is incorporating domain knowledge. In this research no medical or claims knowledge was leveraged. Domain experts may again be able to decrease the number of features by identifying those that they consider strong predictors of cost or readmission.

In this work we also saw a strong bias in the data toward a 30 day readmission label. The impact of over or under sampling should be investigated in creating less bias models. If a uniform distribution of Y was used to train the models perhaps accuracy would improve as it would remove this bias.

Another potential improvement is to use different base models in chained classification or binary relevance for risk or cost models. In our experiments we used the same base model for each, but it could be different models perform better for the different labels.

REFERENCES

[1] Elizabeth J. Atkinson, Mayo Foundation, and Terry M. Therneau. An introduction to recursive partitioning using the rpart routines. 2014.

[2] Dimitris Bertsimas, Margrét V. Bjarnadóttir, Michael A. Kane, J. Christian Kryder, Rudra Pandey, Santosh Vempala, and Grant Wang. Algorithmic prediction of healthcare costs. *Journal of Operations Research*, 2008.

[3] Healthcare Cost and Utilization Project. Washington state inpatient database. *http://www.hcup-us.ahrq.gov*, 2009–2012.

[4] Bala Deshpande. Reasons why feature selection is important in predictive analytics. *http://www.simafore.com*, 2011.

[5] S. Godbole and S. Sarawagi. Discriminative methods for multi-label classification. *Advances in Knowledge Discovery and Data Mining*, pages 22–30, 2004.

[6] Donzé J., Aujesky D., Williams D., and MD. Schnipper J.L. Potentially avoidable 30-day hospital readmissions in medical patients:derivation and validation of a prediction model. *JAMA Internal Medicine*, pages 173(8):632–638, 2013.

[7] K.Zolfaghar, N. Meadem, A. Teredesai, S. Basu Roy, S.C. Chin, and B.Muckian. Big data solutions for predicting risk-of-readmission for congestive heart failure patients. *BigData'13 Big Data in Bioinformatics and Health Informatics Workshop*, 2013.

[8] James Marquardt, Prabhu Ram, Stacey Newman, Viren Prasad, Deepa Hattarki, David Hazel, Archana Ramesh, Rajagopalan Srinivasan, Martine De Cock, Ankur Teredesai, and Shanu Sushmita. Healthscope: An interactive distributed data mining framework for scalable prediction of healthcare costs. *ICDM*, 2014.

[9] Microsoft. Azure machine learning. *http://azure.microsoft.com*, 2015.

[10] Piotr Romanski and Lars Kotthoff. Fselector. *http://cran.r-project.org*, 2014.

[11] Hunter T., Nelson J., and Birmingham J. Preventing readmissions through comprehensive discharge planning. *Prof Case Manag.*, pages 56–63, 2013.

APPENDIX A

	Predicted Label								
True Label	**High Cost 90 days**	**Medium Cost 90 days**	**Low Cost 90 days**	**High Cost 60 days**	**Medium Cost 60 days**	**Low Cost 60 days**	**High Cost 30 days**	**Medium Cost 30 days**	**Low Cost 30 days**
High Cost 90 days	**0**	0	0	0	0	0	89.50%	9.25%	1.25%
Medium Cost 90 days	0	**0**	0	0	0	0	80.81%	14.86%	4.32%
Low Cost 90 days	0	0	**0**	0	0	0	76.98%	19.42%	3.60%
High Cost 60 days	0	0	0	**0**	0	0	93.40%	5.61%	0.99%
Medium Cost 60 days	0	0	0	0	**0**	0	82.37%	14.39%	3.24%
Low Cost 60 days	0	0	0	0	0	**0**	79.82%	15.79%	14.39%
High Cost 30 days	0	0	0	0	0	0	**91%**	8.10%	2.20%
Medium Cost 30 days	0	0	0	0	0	0	85.17%	**11.91%**	2.92%
Low Cost 30 days	0	0	0	0	0	0	78.48%	17.64%	**3.89%**

Table IX

Confusion Matrix using 10-fold cross validation for the binary relevance method and an SVM base model.

Appendix B

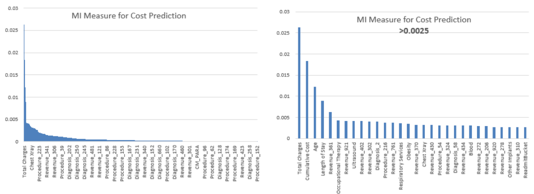

Figure 1. Sorted attributes by MI measure for cost prediction

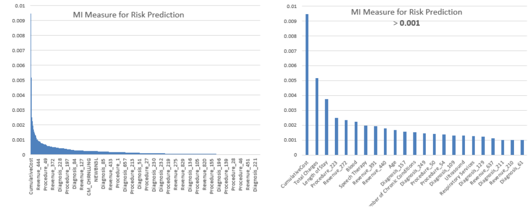

Figure 2. Sorted attributes by MI measure for risk prediction

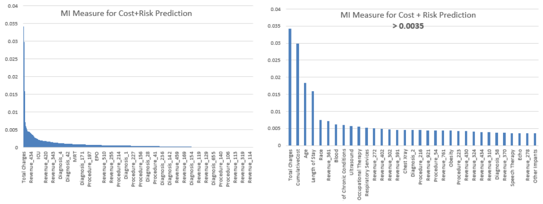

Figure 3. Sorted attributes by MI measure for risk+cost prediction

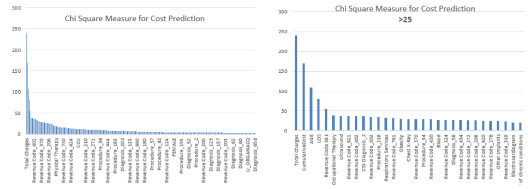

Figure 4. Sorted attributes by CHI measure for cost prediction

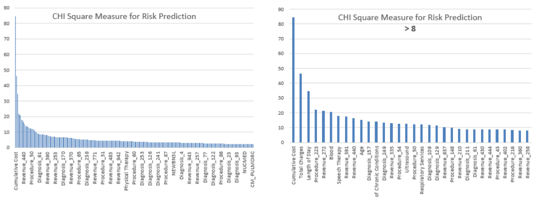

Figure 5. Sorted attributes by CHI measure for risk prediction

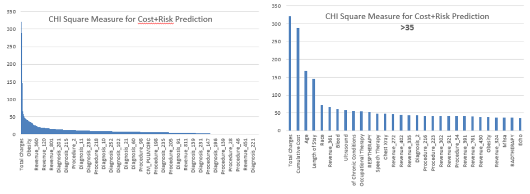

Figure 6. Sorted attributes by CHI measure for risk+cost prediction

APPENDIX C

Measure	Features for Cost
Chi Square greater than 20	Total Charges, Cumulative Cost, Age, Length of Stay, Revenue361, Occupational Therapy, Ultrasound, Revenue921, Revenue402, Diagnosis2, Revenue303, Procedure216, Respiratory Services, Revenue761, Obesity, Chest Xray, Revenue370, Procedure54, Revenue430, Revenue324, Revenue310, Diagnosis58, Revenue434, Revenue272, Revenue206, Revenue920, Revenue278, Other Implants, Electrocardiogram, Blood, Number of Chronic Conditions
Mutual Information greater than 0.0025	Total Charges, Cumulative Cost, Age, Length of Stay, Revenue361, Occupational Therapy, Revenue921, Ultrasound, Revenue402, Revenue302, Diagnosis2, Procedure216, Revenue761, Respiratory Service, Obesity, Revenue370, Chest Xray, Revenue430, Procedure54, revenue324, Diagnosis58, Revenue434, Blood, Revenue272, Revenue206, Revenue920, Revenue278, Other Implants

Table X
FEATURES FROM FILTER-BASE FEATURE SELECTION FOR COST PREDICTION

Measure	Features for Risk
Chi Square greater than 8	Total Charges, Cumulative Cost, Length of Stay, Procedure223 , Revenue272, Blood, Speech Therapy, Revenue391, Revenue440, Age, Diagnosis157, Number of Chronic Conditions, Diagnosis 249, Revenue335, Procedure54, Ultrasound, Procedure50, Respiratory Services, Diagnosis109, Diagnosis129, Revenue637, Procedure148, Revenue210, Diagnosis211, Diagnosis61, Revenue430, Revenue444, Procedure43, Revenue402, Procedure216, Revenue360, Revenue258
Mutual Information greater than 0.001	Total Charges, Cumulative Cost, Length of Stay, Procedure223 , Revenue272, Speech Therapy, Revenue391, Revenue440, Age, Blood, Diagnosis157, Number of Chronic Conditions,Diagnosis 249, Procedure50, Procedure54, Diagnosis109, Ultrasound, Respiratory Services, Diagnosis129, Revenue637, Diagnosis211 Revenue210, Diagnosis61

Table XI
FEATURES FROM FILTER-BASE FEATURE SELECTION FOR RISK PREDICTION

Measure	Features for Cost+Risk
Chi Square greater than 35	Total Charges, Cumulative Cost, Age, Length of Stay, Race, Number of Chronic Conditions, Blood, Respiratory Services, Revenue361, Occupational Therapy, Ultrasound, Speech Therapy, Chest Xray, Revenue272, Revenue402, Revenue335, Diagnosis2, Procedure216, Revenue223, Revenue302, Revenue921, Procedure54, Revenue391, Revenue761, Revenue430, Obesity, Revenue324, RadioTherapy, Echo, Revenue434
Mutual Information greater than 0.0035	Total Charges, Cumulative Cost, Age, Length of Stay , Race, Number of Chronic Conditions, Revenue361, Ultrasound, Blood, Respiratory Services, Occupational Therapy, Revenue272, Revenue402, Revenue302, revenue391, Chest Xray, Diagnosis2, Procedure216, Revenue921, Procedure54, Revenue761, Obesity, Procedure223, Revenue430, Revenue324, Revenue434, Revenue310, Diagnosis58, Revenue370, Speech Therapy, Revenue278, Echo, Other Implants

Table XII
FEATURES FROM FILTER-BASE FEATURE SELECTION FOR COST+RISK

Measure	Top 20 Predictors of Cost
Information Gain	Total Charges, Cumulative Cost, Age, Length of Stay, Revenue361, Occupational Therapy, Revenue921, Ultrasound, Revenue402, Revenue302, Diagnosis2, Procedure216, Race, Respiratory Services, Chest XRay, Revenue761, Obesity, Revenue370, Revenue430, Procedure54
Gain Ratio	Procedure212, Revenue180, Revenue404, Diagnosis243, Procedure103, Procedure144, Procedure32, Procedure187, Revenue331, Revenue280, Revenue729, Revenue542, Revenue949, Revenue946, Diagnosis222 Revenue459, Diagnosis92, Diagnosis170, Procedure2, Diagnosis76
Symmetrical Uncertainty	Total Charges, Cumulative Cost, Age, Length of Stay, Revenue361, Diagnosis2, Revenue921, Procedure216, Revenue302, Revenue761, Revenue402, Occupational Therapy, Ultrasound, Obesity, Revenue370, Procedure54, Chest XRay Revenue920, Revenue310, Revenue430,

Table XIII
FEATURES FROM WRAPPER-BASE FEATURE SELECTION FOR COST PREDICTION

Measure	Top 20 Predictors of Readmission Time
Information Gain	Cumulative Cost, Procedure223, Blood, Speech Therapy, Revenue272, Revenue391, Radio Therapy, Revenue440, Ultrasound, Procedure50, Diagnosis157, Diagnosis249, Race Diagnosis109,Procedure54, Diagnosis129, Respiratory Services, Procedure4, Procedure43, Diagnosis61
Gain Ratio	Diagnosis34, Procedure103, Procedure31, Procedure56, Procedure187, Revenue542, Revenue946, Revenue684, Procedure148, Diagnosis170, Procedure154, Procedure185, Revenue330, Diagnosis172, Diagnosis216, Diagnosis219, Procedure124, Procedure52, Procedure175, Procedure87
Symmetrical Uncertainty	Cumulative Cost, Procedure223, Radio Therapy, Procedure50, Blood, Speech Therapy, Revenue391, Diagnosis157, Revenue272, Diagnosis249, Revenue440, Diagnosis109, Procedure4, Procedure43, Diagnosis61, Diagnosis129, Revenue335, Ultrasound, Procedure54, Procedure44

Table XIV
FEATURES FROM WRAPPER-BASE FEATURE SELECTION FOR PREDICTING TIME OF READMISSION

Measure	Top 20 Predictors of Time and Cost
Information Gain	Total Charges, Cumulative Cost, Age, Length of Stay, Race, Revenue61, Blood, Ultrasound, Procedure223, Respiratory Services, Occupational Therapy, Revenue272, Revenue402, Revenue391, Revenue302, Chest XRay, Procedure216, Diagnosis2, Revenue921, Procedure54
Gain Ratio	Procedure103, Procedure187, Revenue542 Revenue946, Procedure144, Diagnosis170, Diagnosis34, Procedure31, Procedure56, Revenue684, Diagnosis216, Procedure124 Procedure195, Revenue471, Revenue335, Diagnosis172, Diagnosis219, Procedure52, Procedure175, Procedure87
Symmetrical Uncertainty	Total Charges, Cumulative Cost, Age, Length of Stay, Revenue361, Race, Blood, Ultrasound, Procedure223, Occupational Therapy, Diagnosis2, Revenue391, Procedure216, Respiratory Services, Revenue302, Revenue402, Revenue272, radio Therapy, Procedure54, Revenue761

Table XV
FEATURES FROM WRAPPER-BASE FEATURE SELECTION FOR PREDICTING COST AND TIME OF READMISSION AT THE SAME TIME