Diabetes patient Readmission Prediction using Big Data Analytic tools

Xing Yifan
Jai Sharma

Abstract

As diabetes patient readmission rate is becoming one of the major concerns for many national hospitals in the U.S. It is of great significance to study the possible causes and risks of diabetes patients' readmission. In this project, the relationship between diabetes and the various patient attributes are examined. Further, several prediction models are built base on different sets of attributes of the patient. They include a set of numerical attributes including number of outpatient visits, number of emergency visits and time spent in hospital etc., a set of categorical attributes including where the patient come from, what type of admission the encounter faced and where the patient is sent to after discharge and finally, a bag of administration drugs that the patient took/did not take. Possible insights of the study will include a ranking of significance of these attributes and certain recommendations based on that for the hospitals to consider.

Table of Contents

1. Introduction and Background

After a paper (Biomed Research International, 2014.) appeared on HbA1c correlation on diabetic readmission, the result generated by the paper and the data and statistics used are of great influence and the test is then used widely. In this project, rather than verifying the result generated by the paper, we would like to examine diabetes patient readmission rate using a different approach- by classification. Furthermore, considering the other data in the set, we will be analyzing the set of drug administration data, which was not analyzed by the original paper. We would also like to examine some of the categorical fields in about a diabetes patient such as admission source, admission type and discharge position and whether these fields have an impact on the readmission rate or not. Finally, we would like to draw a correlation between several of the continuous variables in our dataset, such as age, number of procedures, number of medications, and time spent in the hospital and the readmission probability.

The database used in this study is from the Health Facts database (Cerner Corporation, Kansas City, MO), a national data warehouse that collects comprehensive clinical records across hospitals throughout the United States. Health Facts is a voluntary program offered to organizations which use the Cerner Electronic Health Record System. The database contains data systematically collected from participating institutions electronic medical records and includes encounter data (emergency, outpatient, and inpatient), provider specialty, demographics (age, sex, and race), diagnoses and in-hospital procedures documented by ICD-9-CM codes, laboratory data, pharmacy data, in-hospital mortality, and hospital characteristics. All data were identified in compliance with the Health Insurance Portability and Accountability Act of 1996 before being provided to the investigators. Continuity of patient encounters within the same health system (EHR system) is preserved.

The Health Facts data the study used was an extract representing 10 years (1999–2008) of clinical care at 130 hospitals and integrated delivery networks throughout the United States: Midwest (18 hospitals), Northeast (58), South (28), and West (16). Most of the hospitals (78) have bed size between 100 and 499, 38 hospitals have bed size less than 100, and bed size of 14 hospitals is greater than 500.

The database consists of 41 tables in a fact-dimension schema and a total of 117 features. The database includes 74,036,643 unique encounters (visits) that correspond to 17,880,231 unique patients and 2,889,571 providers. Because this data represents integrated delivery network health systems in addition to stand-alone hospitals, the data contains both inpatient and outpatient data, including emergency department, for the same group of patients. However, data from out-of-network providers is not captured.

The dataset was created in two steps. First, encounters of interest were extracted from the database with 55 attributes. This dataset is available as a Supplementary Material available online at http://dx.doi.org/10.1155/2014/781670.

The research article did the following evaluations:

Title of the article: Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records (Published 3 April 2014)

- The database had 70,000 inpatient diabetes encounters (17 million unique patients) with sufficient detail for analysis.

- They performed Multivariable logistic regression to fit relationship b/w: measurement of HbA1c and early readmission.

- They concluded with the Result that Probability of Readmission & HbA1c measurement depends on the Primary Diagnosis.

Scope- The Questions Our Model Answered

- Do specific drugs or combinations of drugs indicate likelihood of readmission?
- How does age factor into the probability of readmission and/or outcomes?
- Does the number of procedures, medication, and lab procedures correlate with either the readmission probability or the likelihood that the HbA1c test is performed?
- Do the categorical fields of Admission Source, Admission Type and Discharge_disposition have a significant impact on the readmission prediction?

2. Data description

Data Set Information:
The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes. Information was extracted from the database for encounters that satisfied the following criteria.
(1) It is an inpatient encounter (a hospital admission).
(2) It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.
(3) The length of stay was at least 1 day and at most 14 days.
(4) Laboratory tests were performed during the encounter.
(5) Medications were administered during the encounter.
The data contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc.

The following lists some significant features.

a. ‘Bag-of-Drugs’: Range of all drugs range is 24 fields. Steady for administered and no for not. has the form of a sparse bag-of-words
b. ‘HbA1c’ test result: ranges from test is not performed, acceptable range, and abnormal

c. Primary Diagnosis: indexed by ICD and categorical, values can be coupled and reduced (probably need expansion)
d. Age: ranges from the values (0-10) to (90-100), values can be normalized
e. Number of procedure: continuous values
f. Number of medications: continuous values
g. Number of lab procedures: continuous values
h. Time spent in hospital: continuous but small range
i. Categorical Data of Admission Source, Admission Type and Discharge_disposition

repaglinid	nateglinid	chlorprop	glimepiric	acetohex	glipizide	glyburide	tolbutami	pioglitazo	rosiglitaz	acarbose	miglitol	troglitazo	tolazamid	examide	citoglipto	insulin	glyburide	glipizide-	glimepiric	metformi	metforr
No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Up	No	No	No	No	No
No	No	No	No	No	Steady	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Up	No	No	No	No	No
No	No	No	No	No	Steady	No	No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No
No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No
No	No	No	Steady	No	No	No	No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No
No	No	No	No	No	No	Steady	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
No	No	No	No	No	Steady	No	No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No
No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No	No	Steady	No	No	No	No	No
No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No
No	No	No	No	No	No	Up	No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No
Up	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Down	No	No	No	No	No
No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No
No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No
No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Up	No	No	No	No	No
No	No	No	No	No	Steady	No	No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No
No	No	No	No	No	No	Steady	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No
No	No	No	No	No	No	Down	No	No	No	No	No	No	No	No	No	Steady	No	No	No	No	No

3. Tools/algorithms description

(1) Data Parsing Methodology

Get Label vector

- Primary diagnosis(diag1) is categorized into 9 classes
- HbA1C results into 4 classes
- Readmission into 2 classes

Import the CSV file's data to python np(numpy) arrays

Non-Relevant features were left out as their presence was not affecting the analysis and results much and the field names themselves are pretty intuitive of them needing to be left out namely:

-Race -Gender -Patient_Id etc.

These fields can not be quantified and are of lesser or no relevance for our Prediction Model

Truncate Zero Columns

(2) Data preprocessing

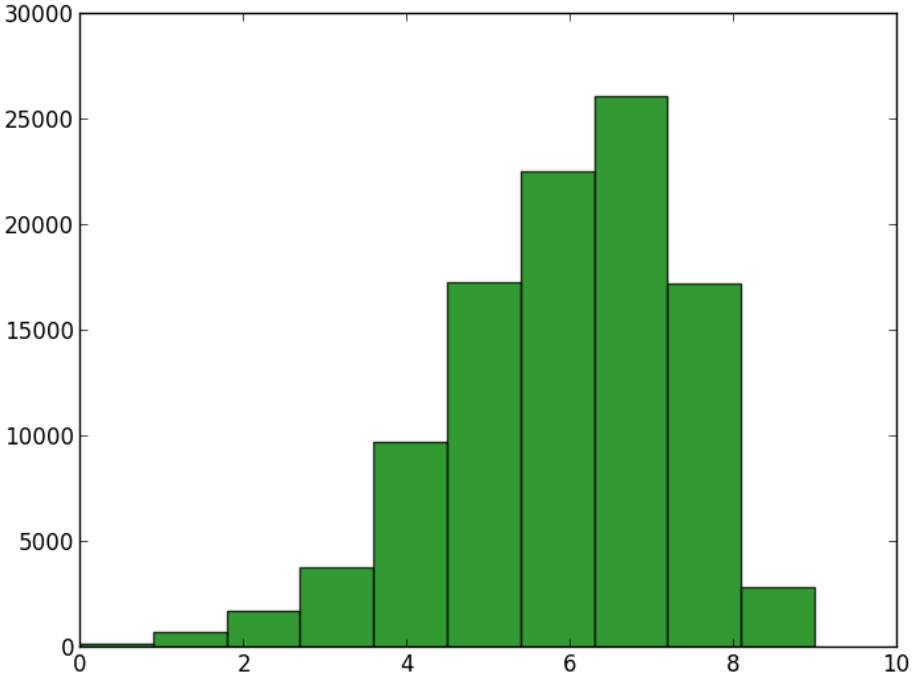
Eliminates Redundant Features

- For Example for Bag of Drugs we eliminated two drugs namely "Examide" & "Citoglipton" as they were not used at all.

Classify Range data to Categorical Data (Encoding Categorical Features)

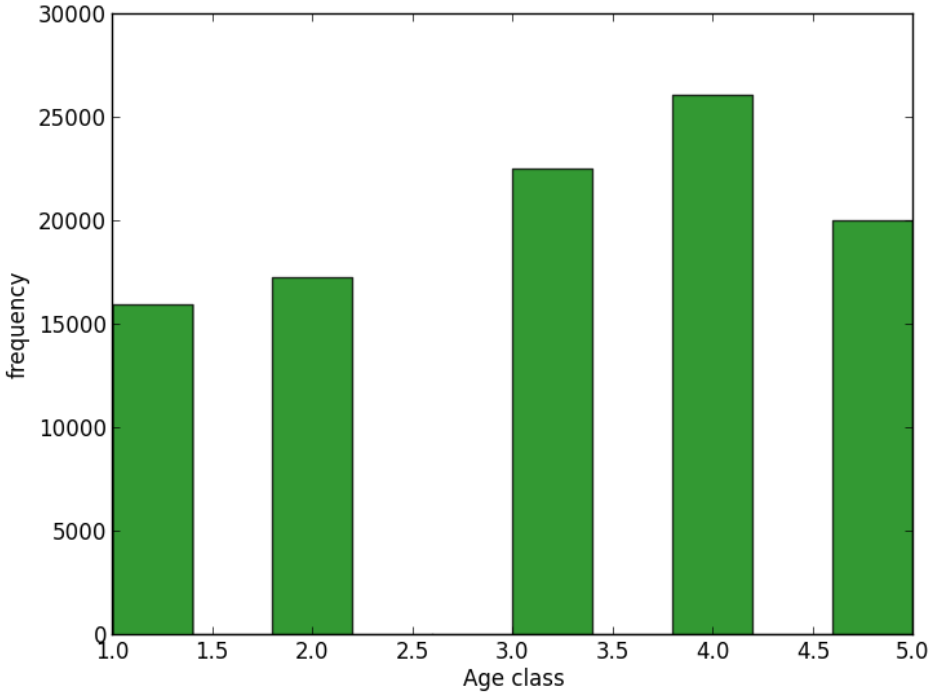
Age Ranges Classified into age classes according to histogram distribution (5 classes in total):

Distribution of age before quantization as given in the original data: (0 means 0-10 years old, 1 means 10-20 years old,...9 means 90-100 etc.)



Distribution of age after quantization is shown below, the 5 age classes are the following:

- Class 1 implying young (0-50 years old),
- Class 2 is median age (50-60)
- Class 3 is moderately old (60- 70)
- Class 4 is old (70-80)
- Class 5 is extremely old (80-100)



Standardize Columns : Datasets might behave badly if the individual feature do not more or less look like standard normally distributed data: Gaussian with zero mean and unit variance..
Normalization Column: In the Normalization process we performed scaling individual samples to have unit norm. This process is useful as if a project plans to quantify the similarity of any pair of samples
TFIDF transformation: For bag of drugs, we also applied TFIDF transformation on the sparse matrix

Missing Data : The model did not need Imputation of data entry with missing value.

Aggregating Label Vector :

Original Readmission Categories :
1. More than 8 days 2. Less than 8 days 3. No Readmission

Our Readmission Categories:
1. Readmission 2. No Readmission

(3) List of Algorithms and Models Used

a. Knearset Neighbor Binary Classification on whether one patient needs to get readmission
   i. Knn on Numerical Data fields as listed below
   using numerical data fields of
   Number of Outpatient visits (counts)
   Number of Inpatient visits (counts)

Number of Emergency visits (counts)
Time spent in hospital (counts)
Number of procedures (counts)
Number of medications: (counts)
Number of lab procedures: (counts)
Age ranges (5 classes)

ii. Knn on Bag of Drugs as mentioned in the data description section

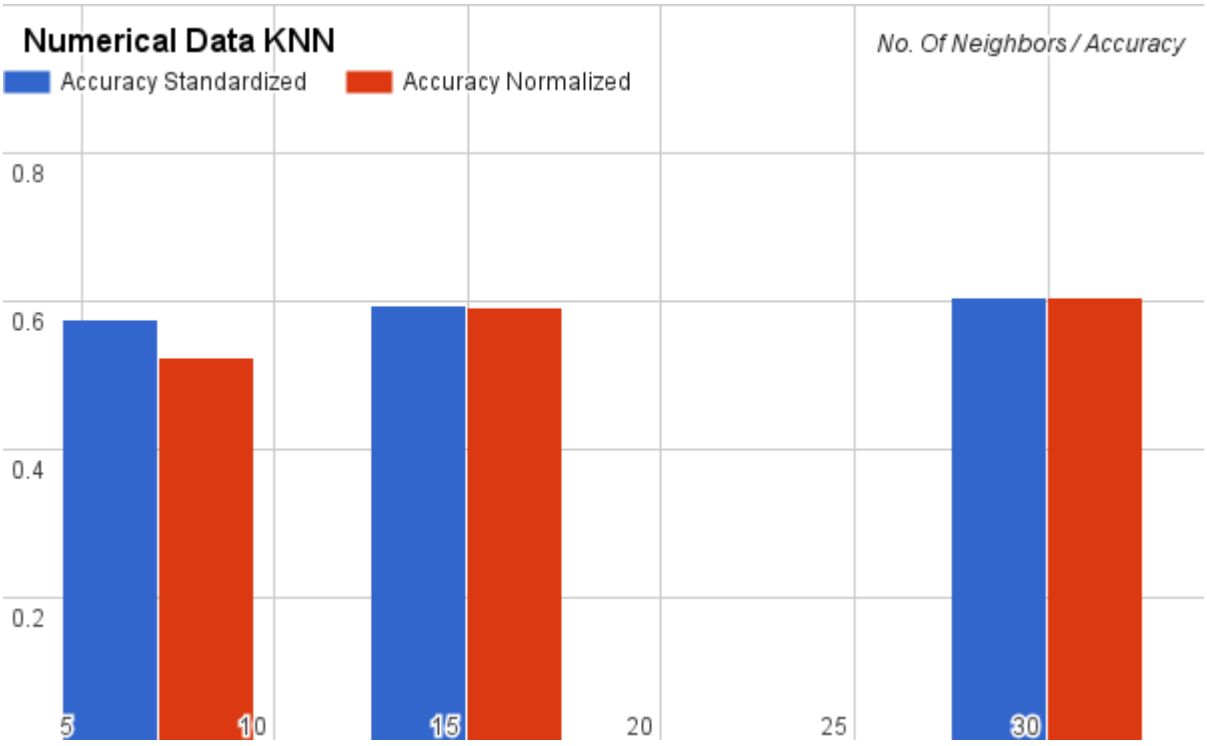
b. Support Vector Machine Binary Classification on readmission rate
    i. SVM using Liblinear core, applied on both the numerical data set and the bag of drugs. It is noticed that Linear_SVM has a high computational efficiency and it is desirable to use it for large data sets with even millions of entries, which is suitable four our case here since we have more that 100,000 samples. Further, when l1 penalty is used in the model, the a sparse matrix of coefficients of the feature variables will be derived and this could be used to indicate the significance of the features in the model.
    ii. SVM using libsvm core, it supports other non linear kernel such as rbf, polynomial and sigmoid rather than just linear kernel and thus, squared hinge loss is by default used. Further it has a high computational complexity of O(n_features * n_samples^3) thus might be slow in our case here.
c. Decision Tree Binary Classification with additional categorical fields on readmission rate
    i. Add the feature of discharge_disposition, admission source , admission type as categorical data to feed the Decision Tree Model
    ii. Apply Decision Tree Model on the Bag of Drugs to compare and contrast the accuracy in prediction as compared to the previous models
d. Ensemble Method in Binary Classification
    i. Built based on the individual prediction models of Knn, SVM and Decision Tree.
    ii. Deploy weighted votes in aggregating and deciding the final prediction label for the testing data set.
e. Feature Selection to Identify Significant fields as predictors to readmission rate that the hospital should pay attention to.
    i. Wrapper Methods:
        1. L1-based feature selection
            a. Selecting Non-Zero Coefficients by tuning the penalty parameter C
            b. Perform L1 in SVM : Non-zero features are significant ones
            c. L1-based tends to select only one feature out of a group a related or correlated feature, this behavior can help eliminate the collinearity between the feature variables, on the other hand, sometimes, patterns based on a combination of the related features may not be discovered.
        2. RFECV- Recursive Feature Elimination and Cross-Validated Selection: Using Cross Validation on the training set, recursively eliminate redundant features until best prediction accuracy is obtained.
    ii. Filter Methods: Variable Ranking

1. Correlation Ranking
2. Chi Square Independence Test between label vector and feature vector to get ranks of significant features
3. Single feature performance ranking using Knn and SVM model

4. Comprehensive experimental results and analysis

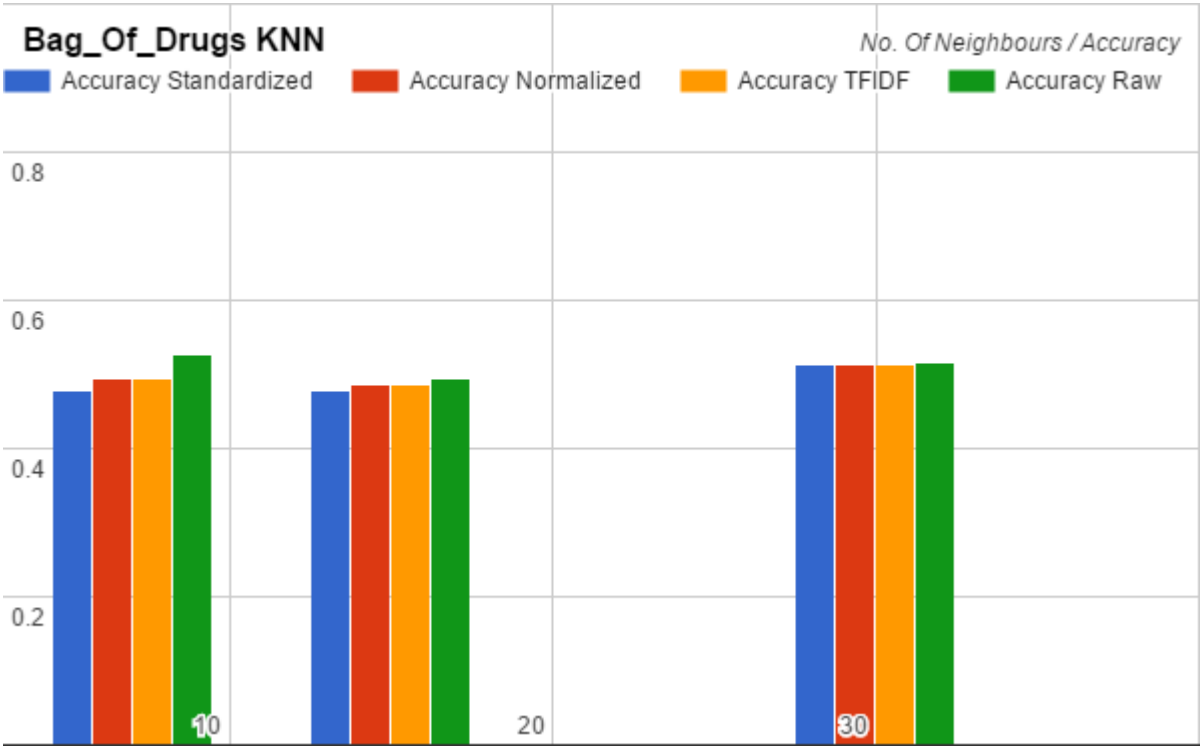
a. Prediction Models - K nearest neighbors

1. Based on the Numerical Data features



The above shows the Knearest Neighbor prediction accuracy based on a 5 fold cross validation on the numerical data. It is observed that as number of neighbors increases, the prediction accuracy increases and in general the standardized numerical data has a higher prediction accuracy than the normalized numerical data for Knn model. The average accuracy in prediction with Knn is around 60%

2. Based on the Data bag of Drugs



Similar to the numerical data case, the accuracy increases as the number of neighbors increases. In addition, the TFIDF and normalized bag of drugs appear to have better accuracy than the standardized bag of drugs. Further, it is observed that Raw bag of drugs appear to be having the best prediction accuracy for Knn model here. The average accuracy is around 53% for raw bag of drugs Knn prediction model in general with 7 neighbors and 5 fold cross validation.

3. With tuned sigma in Gaussian Similarity for weights of the neighbors

$$s(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{d(\mathbf{x}_i, \mathbf{x}_j)^2}{2\sigma^2}\right)$$

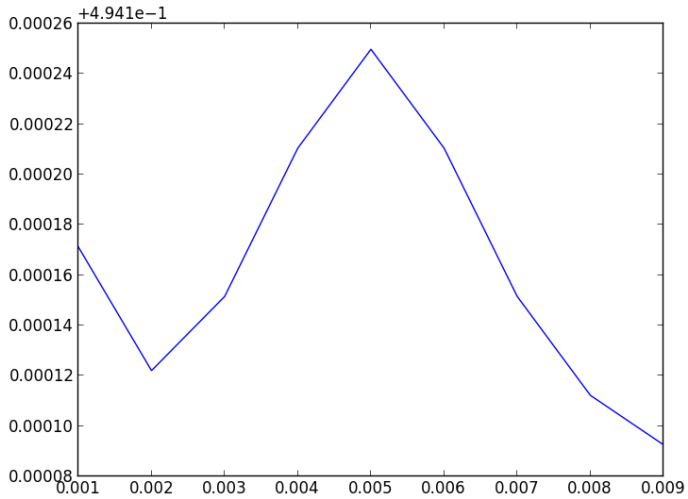
With the idea of incorporating a Gaussian Similarity for weights when predicting the labels from the pool of neighbors rather than the uniform weights, we added the above function and tuned the sigma for Knn model on both data numerical case and bag of drugs case. The distance metric used here is standard Euclidean distance.

(1). Data Numerical case

Tuning sigma for weights for neighbors in the KNN model does not significantly improve the prediction accuracy

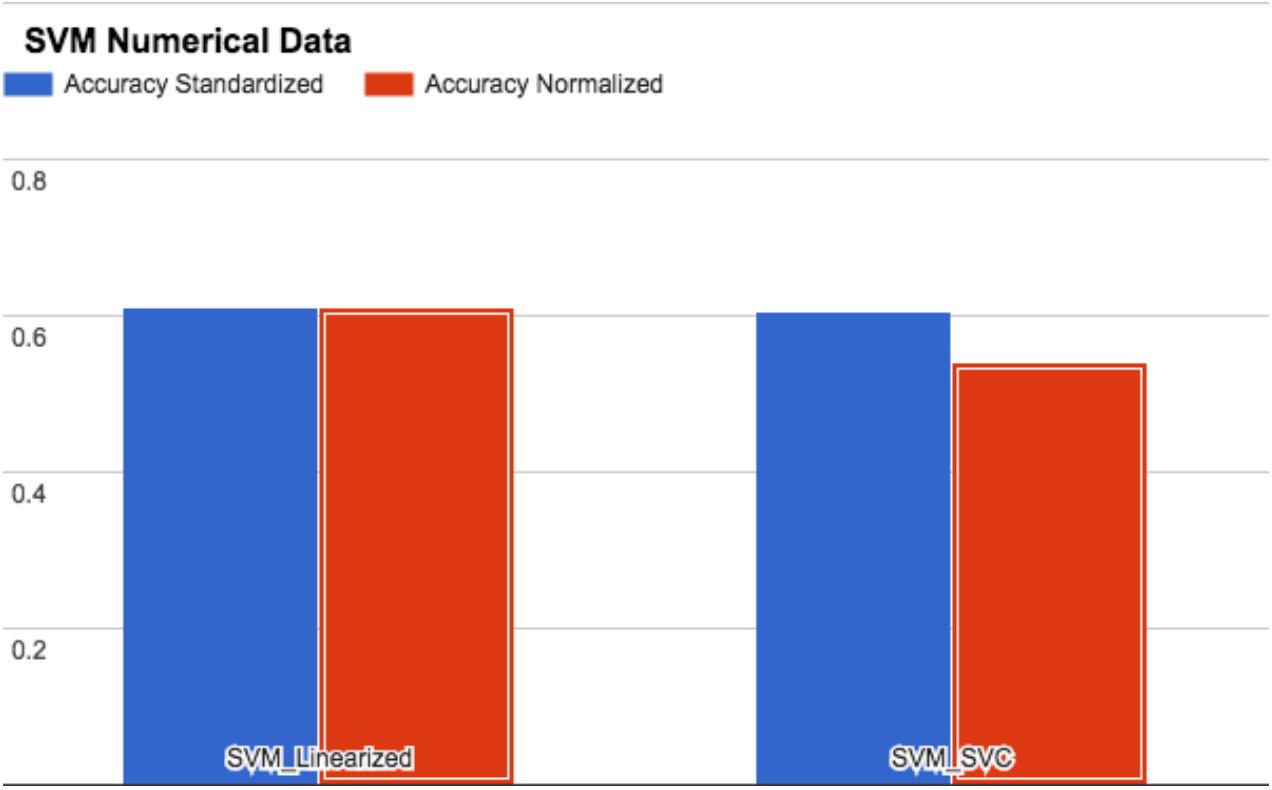
(2). Bag of Drugs case

Tuning of Sigma for neighbors in Knn model provides a 0.02% increase in accuracy as shown in the tuning graph below. The best accuracy is achieved when we tune the sigma to be around 0.005. This is tuning process in done on the normalized bag of drugs thus a 0.005 value for the optimal sigma makes sense since the data matrix is sparse and when normalized, the value of the entries will be small thus yielding a small standard Euclidean distance.



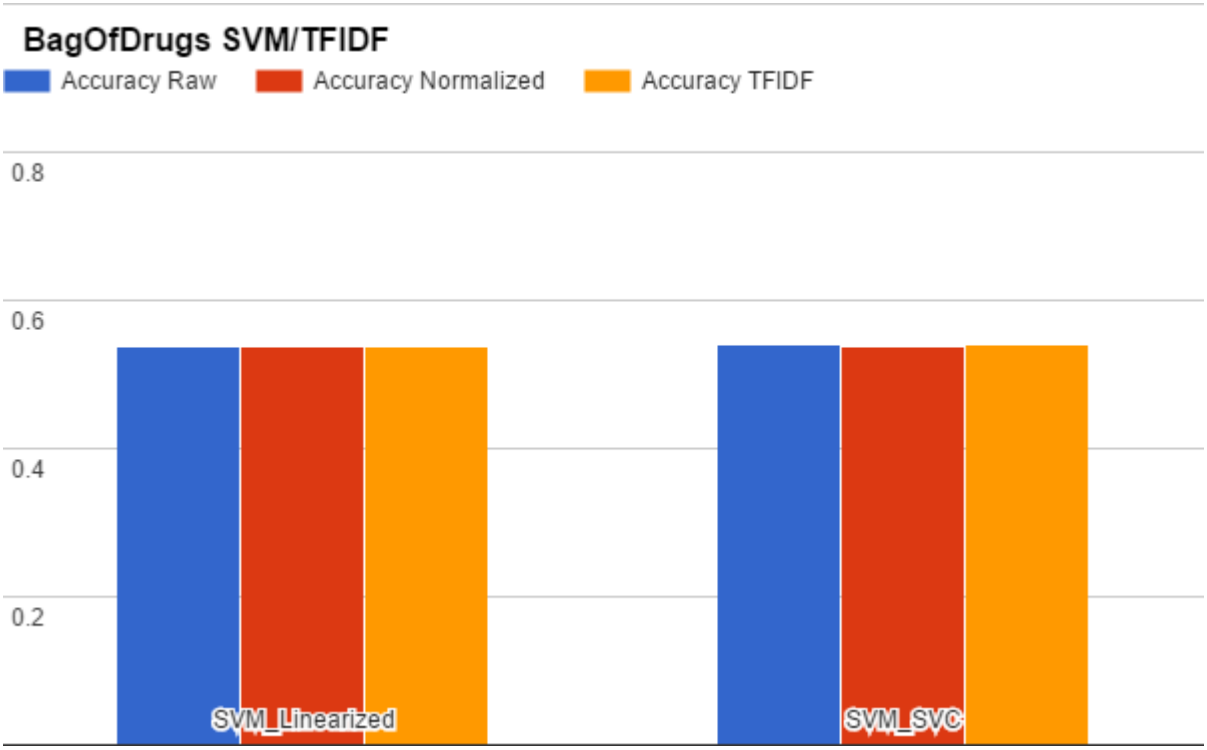
b. Prediction Models - Support Vector Machines

1. SVM prediction accuracy with 5-fold cross validation on the numerical data set.



It is shown that in general SVM linear has a better performance than SVM_SVC which makes use of a default non-linear kernel - 'rbf' kernel, which is essentially the kernel based upon the Gaussian Similarity. Further, standardized numerical data tends to have a better prediction accuracy than normalized numerical data. The average prediction accuracy for linear SVM is around 61% on normalized numerical data.

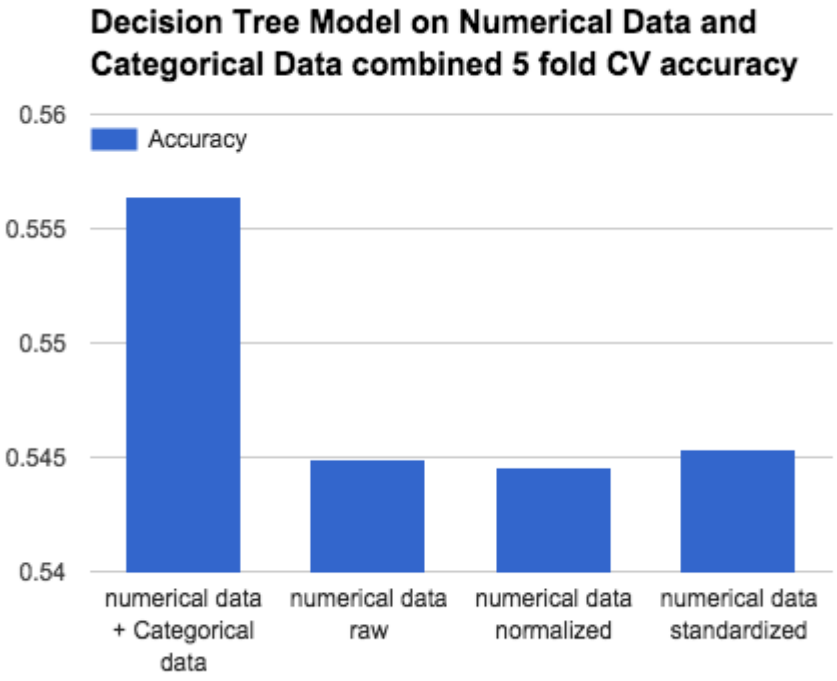
2. SVM prediction accuracy 5 fold cross validation on the Bag of Drugs.



For the case of bag of drugs, the prediction accuracy does not differ much for the Linear SVM and the SVM_SVC based on rbf kernels. Further, the difference in raw, normalized and TFIDF transformed data is minor. The average prediction accuracy for SVM_linear is 53.8% and the average prediction accuracy for SVM_SVC is 54%.

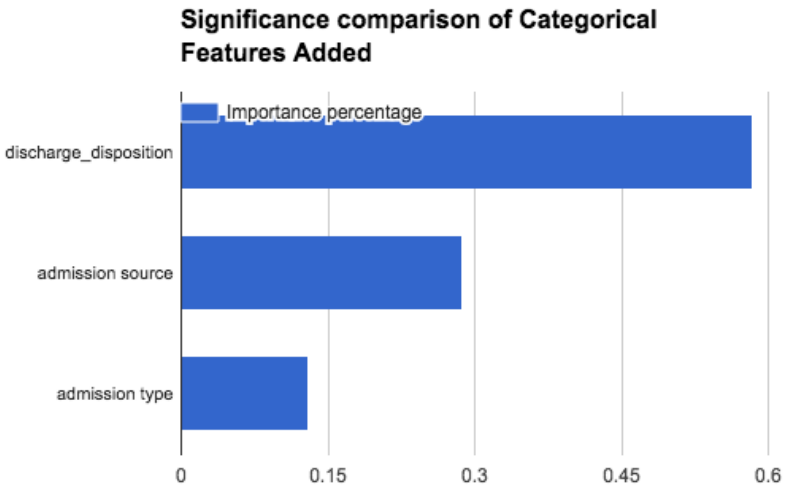
c. Prediction Model- Decision Tree Analysis

1. On Combination of the numerical data and the categorical data



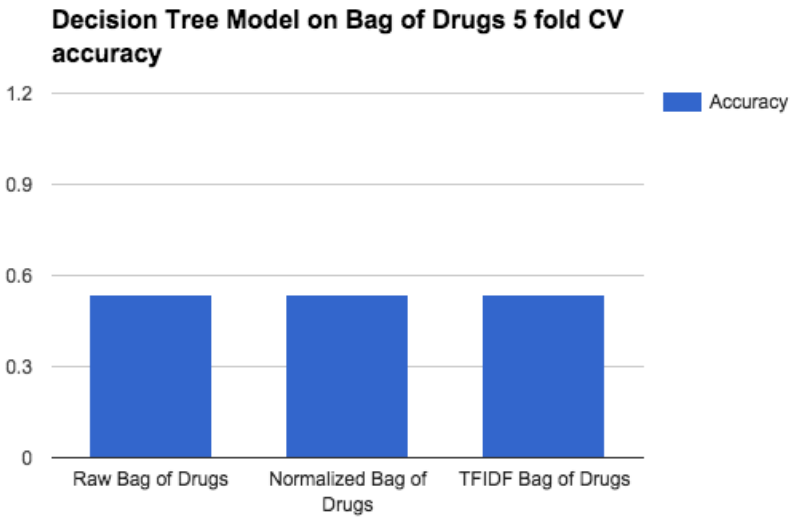
We could see that the prediction accuracy of the decision tree on the numerical data (54.6%

maximum on the standardized numerical data) is not as high as the previous models on the numerical data (Knn yields accuracy around 60% and SVM yields accuracy of around 61%). However, adding the categorical fields of admission source, admission type and discharge_disposition inside the decision tree model which already contains the numerical data does increase the prediction accuracy. Thus, it is concluded that these categorical fields have a significance in helping prediction of the readmission rate as well.



Further, as shown in the chart above, the ranked feature significance of the three categorical features are discharge_disposition > admission source > admission type with the percentage significance of [0.58395725 0.28626858 0.12977416]. This indicates that the hospital might wish to pay attention on these three categorical data in terms of readmission prediction and especially on the attribute of discharge_disposition, ie, where the patient leave for after he/she is discharged.

2. On bag of drugs



It is observed that the average accuracy is around 53.7%. This accuracy is consistent through Raw, Normalized and TFIDF transformed bag of drugs. This accuracy is smaller than the prediction accuracy of the Support Vector Machine but the accuracy is higher than

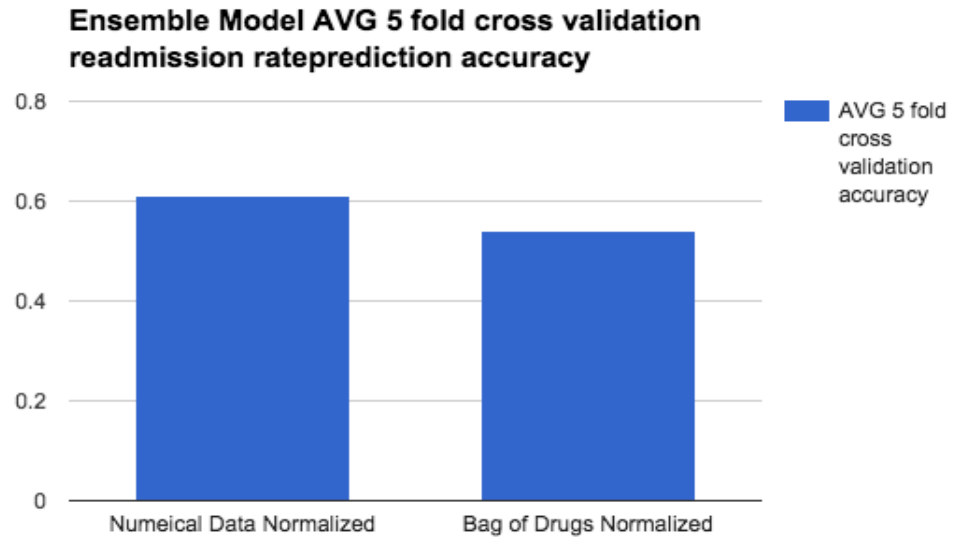
the Knn model.

d. Ensemble Classification Model on readmission rate

It is built based on individual classifiers of Knn, Linear SVM ,SVC SVM and decision tree.

Weighted Votes is used for obtaining the ensemble prediction labels out of the prediction labels from the individual classifiers. In the analysis, the weight of each model's label will be proportional to that model's prediction accuracy based on historical run results. Thus, the higher overall accuracy the individual model has, the more saying it has in the final prediction label.

The ensemble method will eventually have an averaged effects and it aggregates the labels from the relatively weak individual models and aim to produce more sophisticated prediction label in the end. The following chart shows the ensemble model run on the normalized numerical data and the normalized bag of drugs. Overall prediction accuracy of 60.9% and 54.1% are achieved for the two cases. Note that 54.1% accuracy is higher than any of the individual prediction model accuracy could be derived on the normalized bag of drugs data. Thus, the ensemble model does help in generating overall better prediction accuracy by aggregating the votes of the individual prediction models.



e. Feature Selection to identify significant predictors

Feature Selection - Wrapper Methods

1. l1 based feature selection by tuning penalty for slack variable C in the model.

(1). Numerical Data set: by tuning the C, we could control the sparse coefficients matrix generated and the number of features with non-zero coefficients. The following table shows

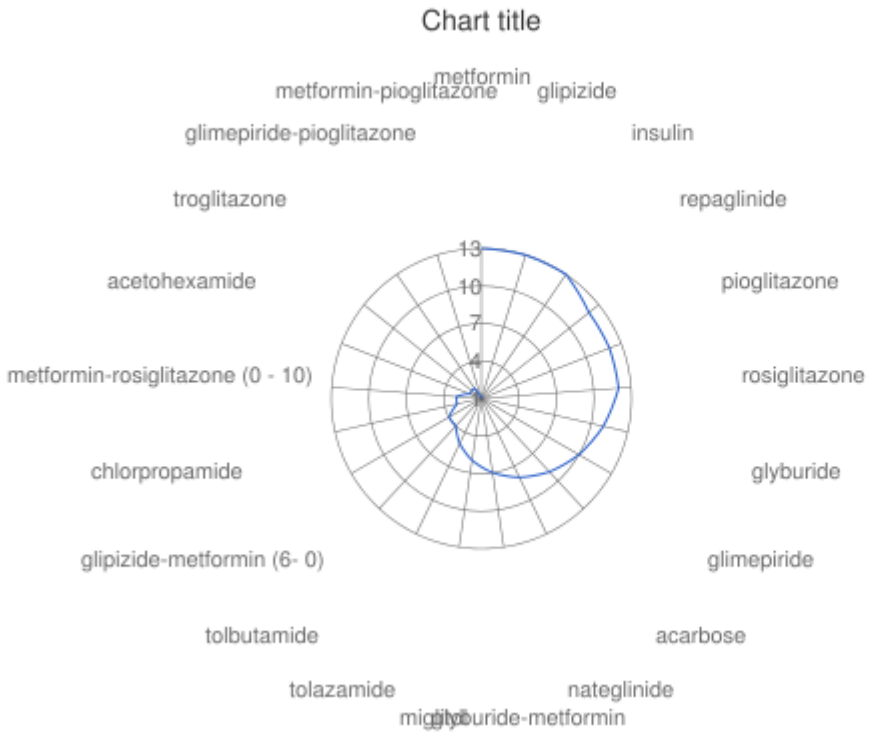
the procedure of tuning C from small to large and incorporating more and more features in the model. Thus, the significances of features are shown in the last column as based on the sequence of being included.

penalty C	Feature Set selected	Features added	Importance of features(decreasing order)	Feature names
numerical_data_penalty = 0.05	[5 6 7]	-	5,6,7	number of Outpatient visits, number of inpatient visits, number of emergency visits
numerical_data_penalty = 0.07	[3 5 6 7]	3	3	Number of procedures
numerical_data_penalty = 0.08	[1 3 5 6 7]	1	1	Time spent in hospital
numerical_data_penalty = 0.11	[0 1 3 5 6 7]	0	0	Age range
numerical_data_penalty = 0.15	[0 1 3 4 5 6 7]	4	4	Number of medications
numerical_data_penalty = 1.0	[0 1 2 3 4 5 6 7]	2	2	Number of lab procedures

Thus, the significance of the numerical features are ranked as: feature [5 6 7] > [3] > [1] > [0] > [4] > [2]

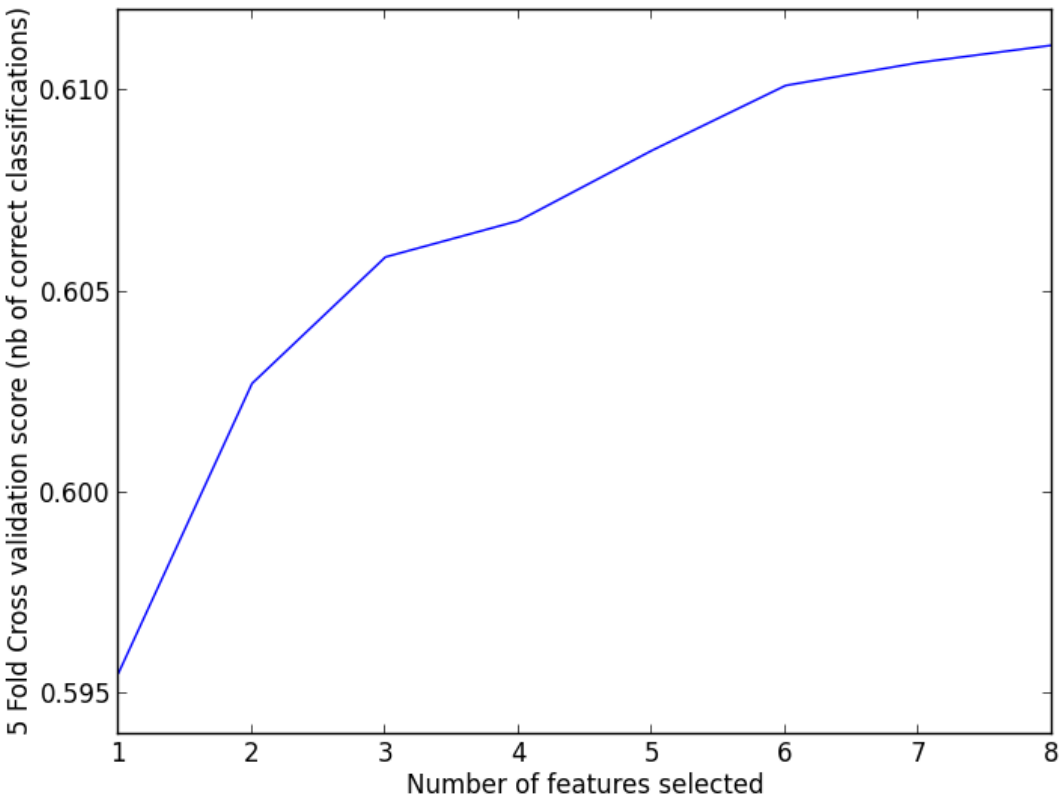
(2) Bag of Drugs
Similar procedure is implemented and the following table shows the rank on the importance of the drugs (the drugs in the list are indexed, for the complete names refer to the spiral chart below)

Penalty C	Feature Set selected	Features added	Importance of features(decreasing order)
0.001	[0 6 15]	-	0,6,15
0.003	[0 1 6 9 10 15]	1,9,10	1,9,10
0.004	[0 1 6 7 9 10 15]	7	7
0.006	[0 1 4 6 7 9 10 15]	4	4
0.01	[0 1 4 6 7 9 10 11 15]	11	11
0.03	[0 1 2 4 6 7 9 10 11 15]	2	2
0.04	[0 1 2 4 6 7 9 10 11 15 16]	16	16
0.05	[0 1 2 4 6 7 9 10 11 12 15 16]	12	12
0.1	[0 1 2 4 6 7 9 10 11 12 14 15 16]	14	14
0.2	[0 1 2 4 6 7 8 9 10 11 12 14 15 16 17]	8,17	8,17
0.3	[0 1 2 3 4 6 7 8 9 10 11 12 14 15 16 17 19]	3,19	3,19
0.5	[0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19]	5,13,18	5,13,18
1	[0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20]	20	20



2. RFECV

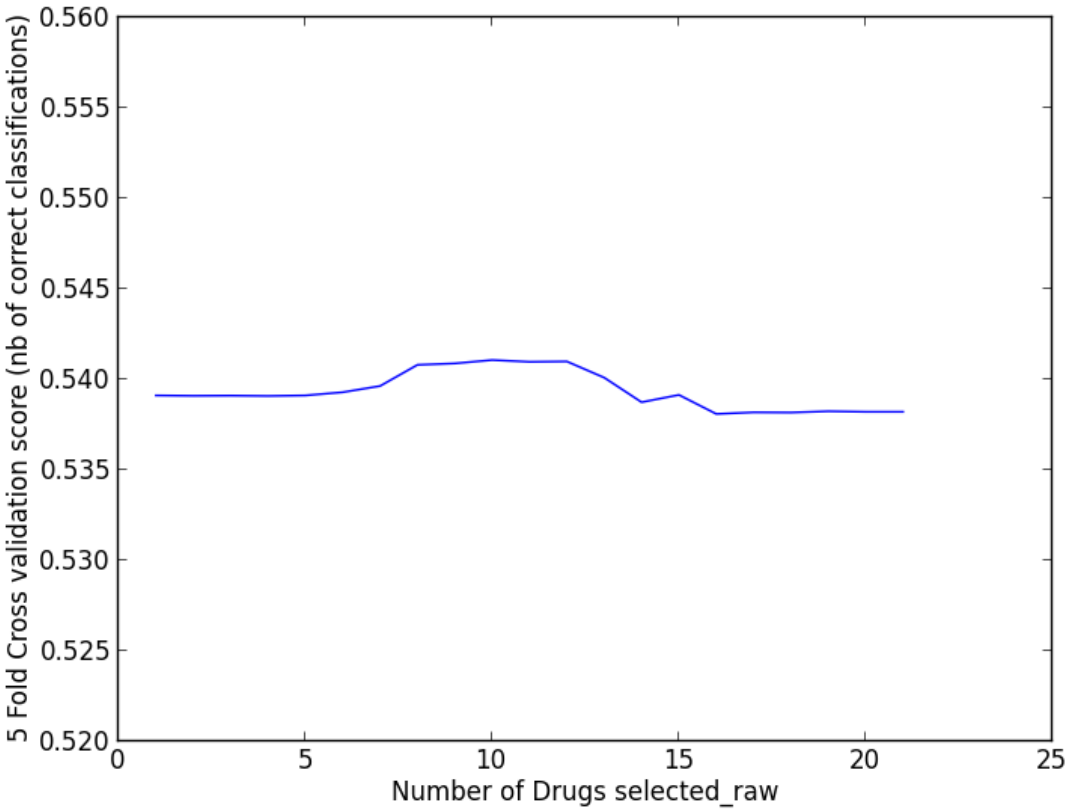
(1). Numerical Data:



It is observed that as we recursively the eliminate the features by 1 each time based on the absolute value of its coefficient in the sparse matrix generated by the l1 linear SVM model, the prediction accuracy just decreases. In another word, by keeping all the numerical

features, the best accuracy could be achieved. This is not surprising as it shows that all the numerical features have certain positive effect in predicting the readmission rate. However, this does not provide us with the ranking of the significant numerical features.

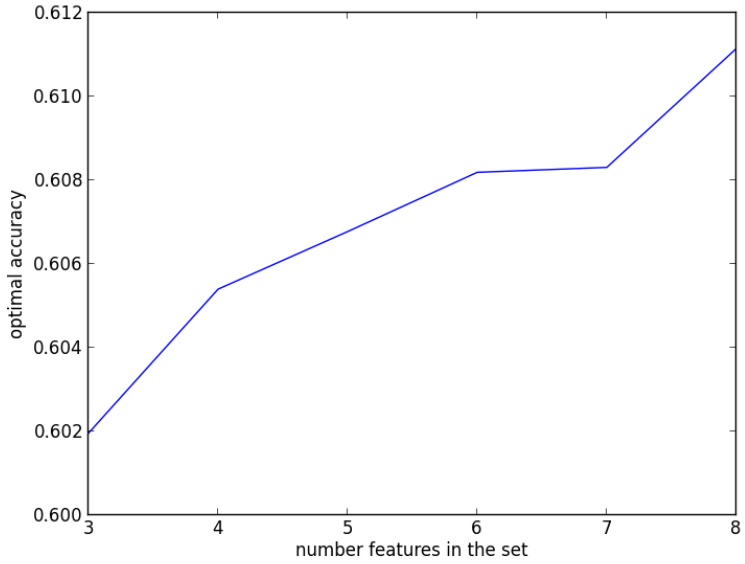
(2). Bag of Drugs



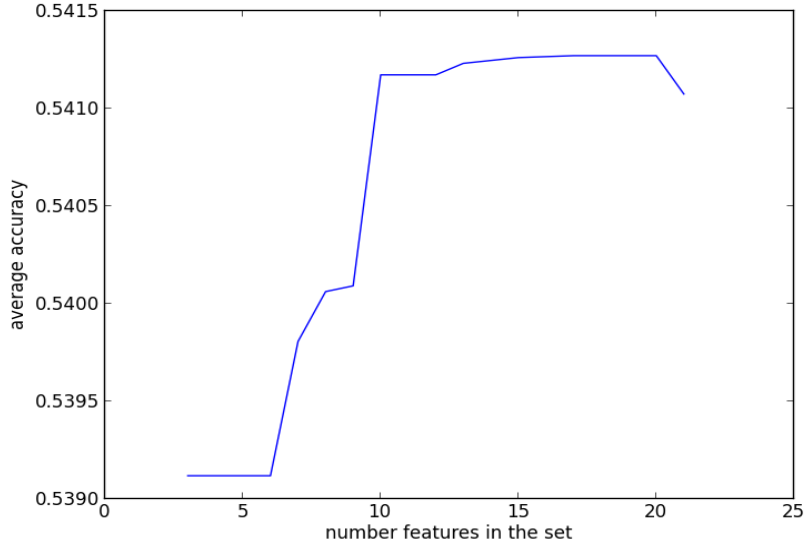
It is observed that as we eliminate the drugs one at a time, the accuracy does not simply decrease, and it is derived that when 10 drugs are kept in the model, the accuracy is optimal. These 10 drugs are of index [5,8,11,12,13,14,17,18,19,20] and are of greater significance. Although it is usually expected that accuracy will increase when the number of drugs selected increases, some fluctuations might happen and since here the range of the fluctuation is pretty small, within 0.3%, the result is acceptable.
Note that this result of set of significant drugs is a bit different from the l1 based features selection approach since it eliminates features one by one by considering only one metric which is the features' absolute value of its coefficients in the sparse matrix and it fail to enumerate all the possible combinations of the drugs. Thus, the result is limited and this set of drug might not be optimal among all combinations but it is a relatively good set which produces decent prediction accuracy.

3. l1 based feature selection + RFECV
By combining the l1 based feature selection and the RFECV approach, we hope to generate a better set of significant features that will optimize the prediction accuracy. The algorithm works as follows, 1) tuning the penalty parameter C in the l1 Linear SVM model to generate a set of feature, call them feature sets 2) out of this feature set tuned by parameter C, run the RFECV algorithm on this feature set to generate the best selected set of feature which

yields the optimal accuracy. This algorithm better enumurates the possible combination of features that will generate better accuracy.
(1). Numerical Data set:



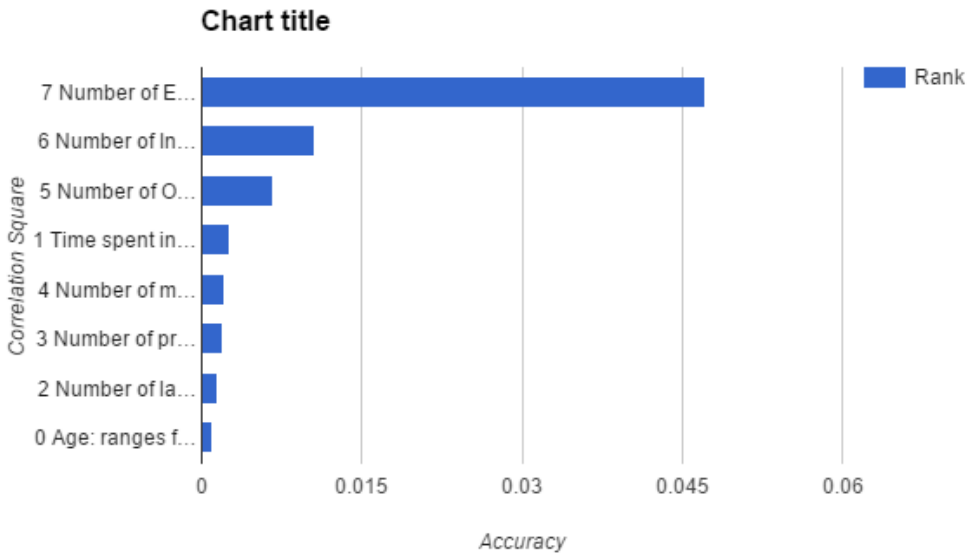
While number of features in the feature set (with non-zero coefficient in the coefficient sparse matrix) increases by tuning up penalty C in the l1 Linear_SVM model, the optimal accuracy increases; the optimal accuracy is still achieved by incorporating all the features inside the model.
(2). Bag of Drugs data set



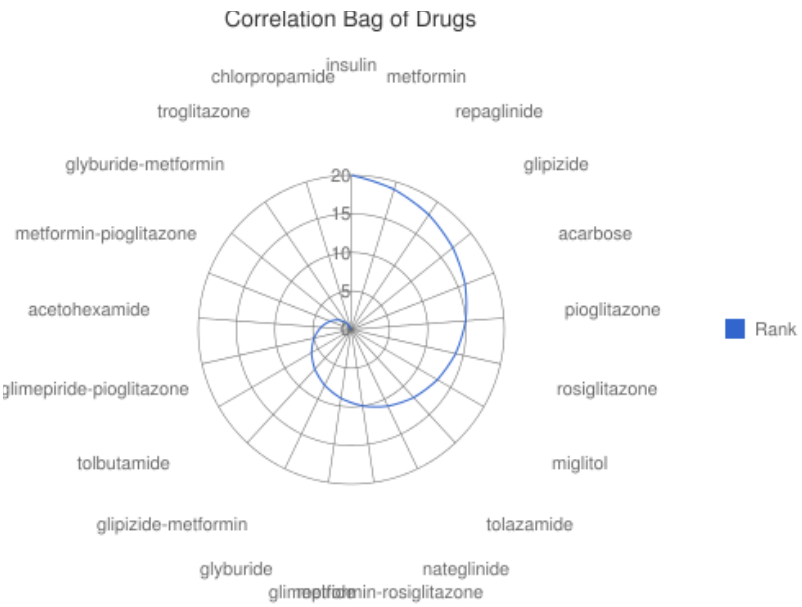
As shown in the graph above, keeping a feature set containing 20 drugs by tuning C penalty and then choose 8 out of this feature set gives the best performance, the drugs selected are of index [0 1 8 11 12 14 15 17] , which are [metformin , repaglinide,tolbutamide , acarbose, miglitol,tolazamide , insulin, glipizide-metformin].

Feature Selection - Filter Method:

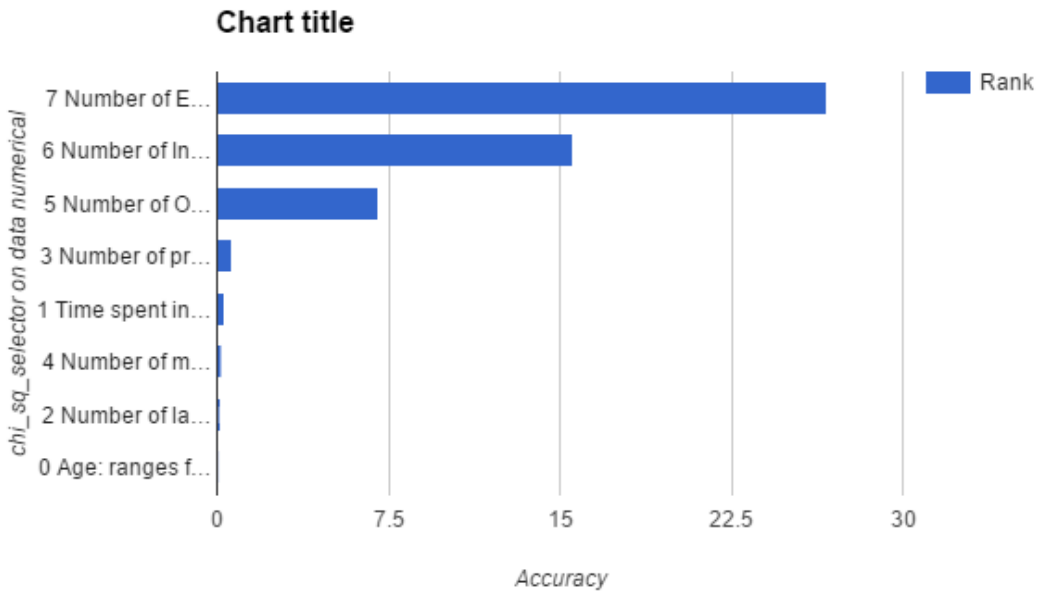
(1). Correlation
(a). Numerical Data: the rank of the significance of features according to correlation is [7 6 5 1 4 3 2 0] in decreasing order. It is observed that features 7,6,5 which are number of emergency visits, inpatient visits and outpatient visits are the 3 most important features in correlation value and the rest features tend to be similar to each other in terms of correlation value.



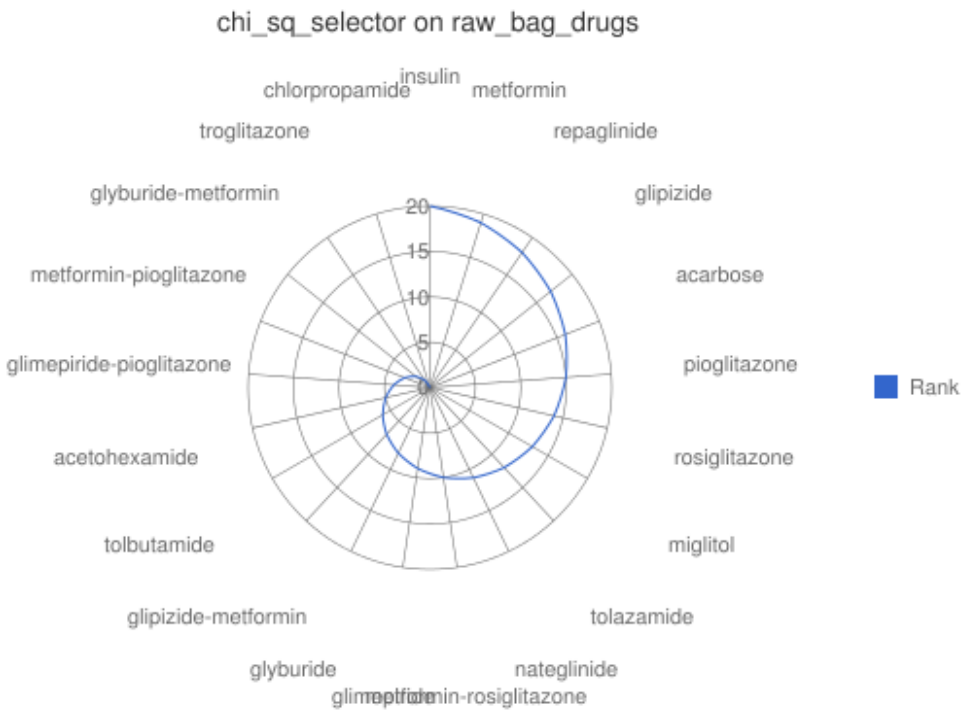
(b). Bag of Drugs: the rank of the drugs on correlation value is shown in the spiral chart:



(2). Chi square independence Test between feature vector and label vector
(a). Numerical Data Set: the rank of the significance of the features based on Chi square independence test is [7 6 5 3 1 4 2 0] in decreasing order. It is observed that features 7,6,5,3 which are number of emergency visits, inpatient visits and outpatient visits and number of procedures are the 4 most significant features in terms of chi square test value.

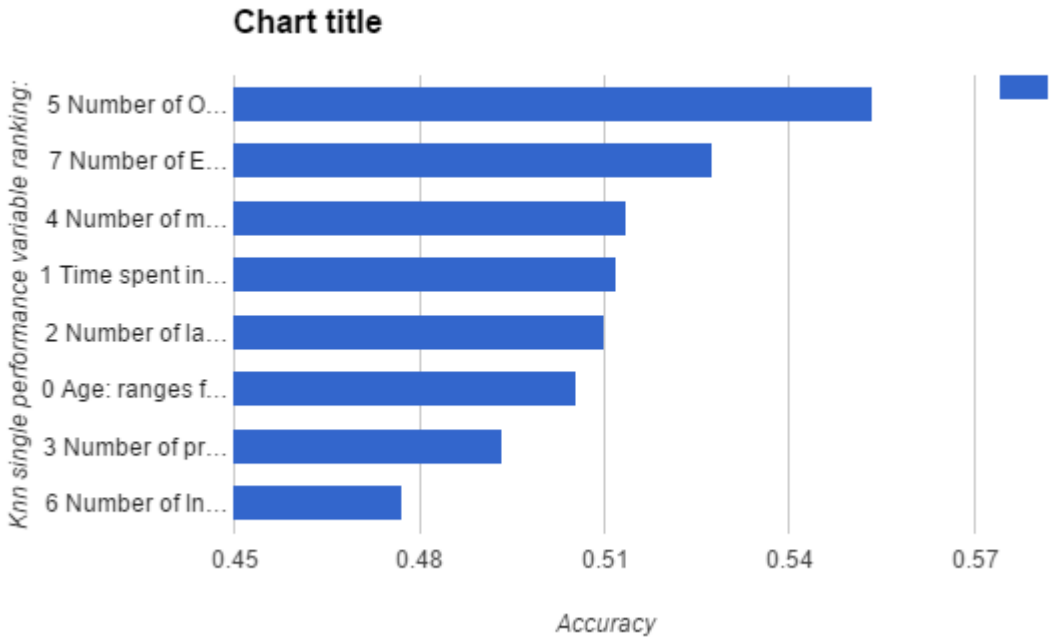


(b). Bag of Drugs: the rank of the drugs on chi_sqaure value is shown in the following spiral chart:

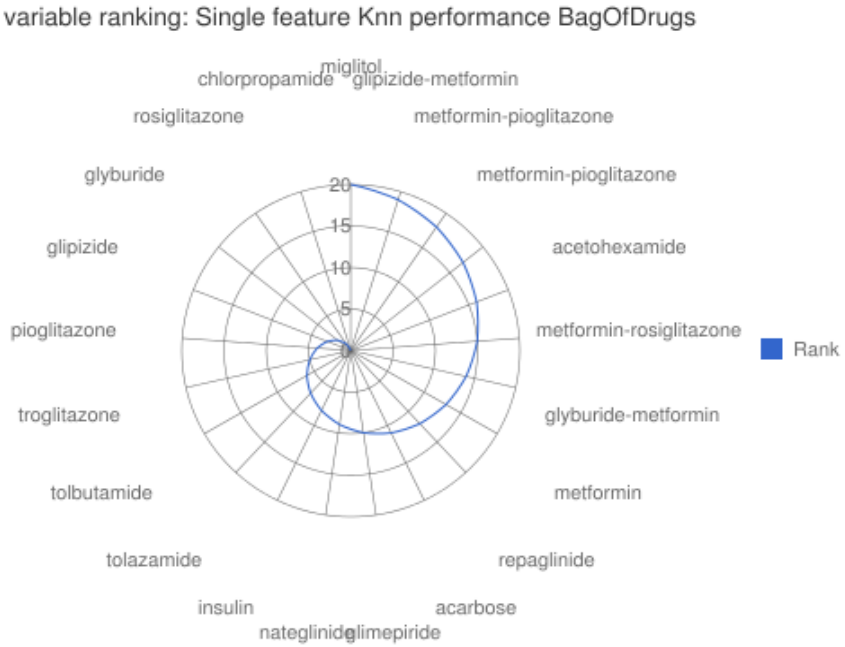


(3). Single Feature Knn performance

(a). Numerical Data Set: the rank of the significance of the features based on single features Knn performance is as shown below along with the prediction accuracy in Knn model with that single feature. It is observed that features 5,7,4,1,2,0 which are number of outpatient visits, emergency visits and medications ,time spent in hospital, number of lab procedures and age range are the 6 most important features in prediction accuracy.

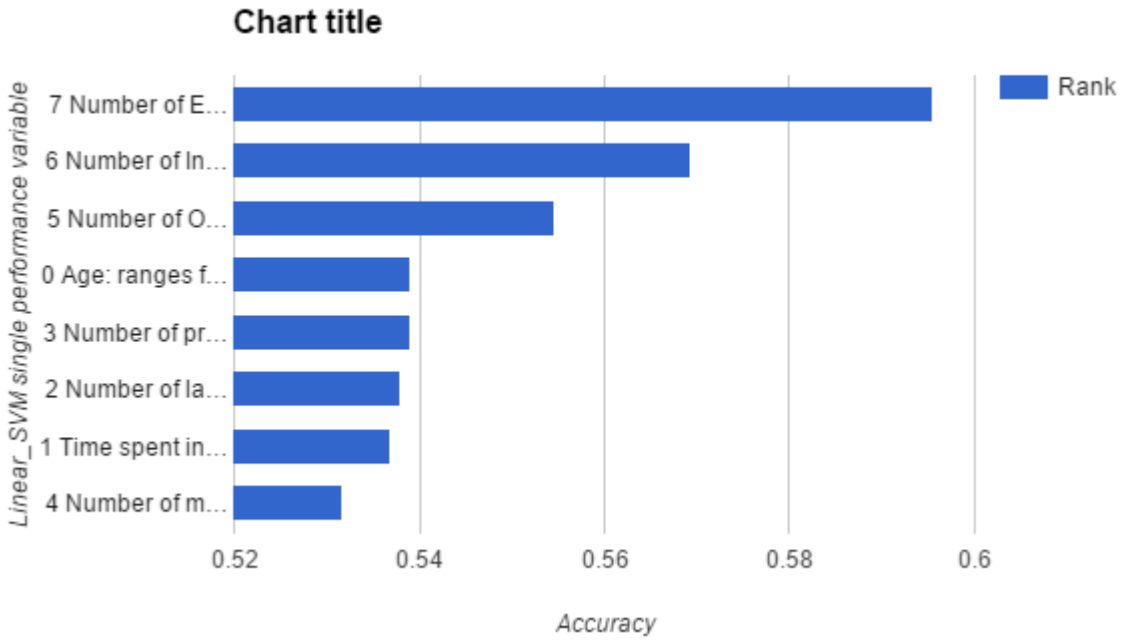


(b). Bag of Drugs: the rank of the drugs on Knn performance is shown in the following spiral chart

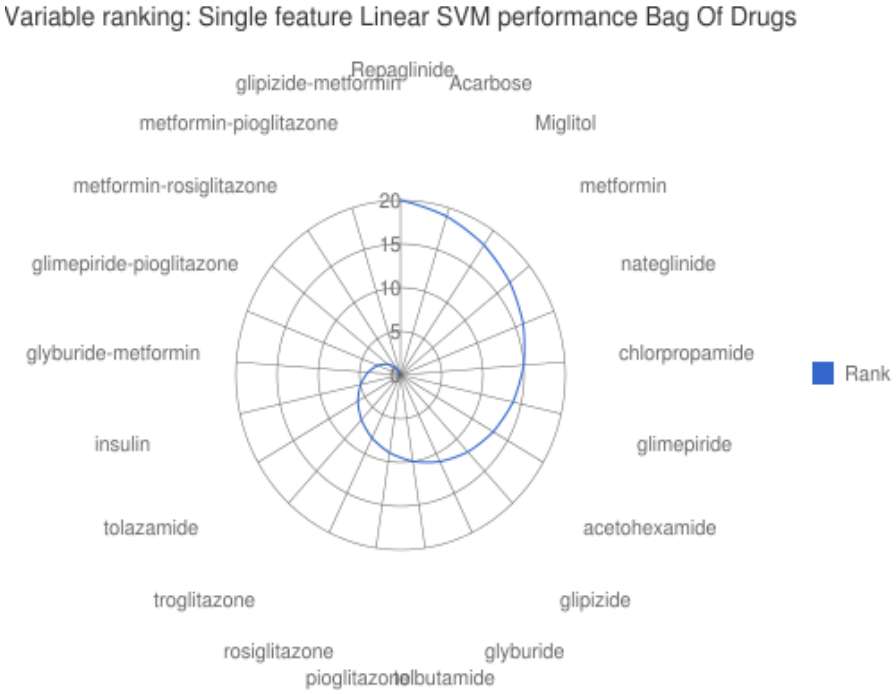


(4). Single feature Linear SVM performance

(a). Numerical Data Set: the rank of the significance of the features based on single features Liner SVM performance is as shown below along with the prediction accuracy in the Linear SVM model with that single feature. It is observed that features 7,6,5,0,3 which are number of emergency visits, inpatient visits and outpatient visits , age range and number of procedures are the 5 most significant features in terms of prediction accuracy.



(b). Bag of Drugs: the rank of the drugs on Linear SVM performance is shown in the following spiral chart



Feature Selection - Result Analysis

1. Numerical Data features:

Most Significant fields(5,6,7)	Moderately Significant fields(1,3,0)	Least Significant fields (4,2)
Number of Outpatient visits Number of Inpatient visits Number of Emergency visits	Time spent in hospital (counts but small range) Number of procedures (counts values) Age : Ranges Quantized into Age Classes (5 classes in total)	Number of medications: counts values Number of lab procedures: counts values

Remark: Since Age range turns out to be a quite important feature as it is categorized in the second most important features set; We conclude that age range does significantly factor into the prediction.

2. Bag of Drugs: By aggregating the information gained from the various variable ranking filter methods and the wrapper methods on the importance of the drugs in prediction accuracy, we compiled the following list of drugs which are categorized into three criteria, the important drugs, moderately important and the least important drugs in terms of prediction accuracy on readmission rate.

Important Drugs' Index	**Important Drugs (In decreasing importance)**
15	insulin
0	metformin
1	repaglinide
6	glipizide
4	glimepiride
7	glyburide
11	acarbose
9	pioglitazone
10	rosiglitazone
12	miglitol
2	nateglinide
Median Important Drugs' Index	**Median Important Drugs in prediction**
17	glipizide-metformin (combination of 6- 0)
8	tolbutamide
14	tolazamide
19	metformin-rosiglitazone (combination of 0-10)

Non-Important Drugs' Index	Non-important Drugs
3	chlorpropamide
5	acetohexamide
13	troglitazone
16	glyburide-metformin (combination of 0-7)
18	glimepiride-pioglitazone (combination of 4-9)
20	metformin-pioglitazone (combination of 0-9)

(Note that for drug with index 16,17,18,19,20, they are actually pills containing two of the previous drugs.)

5. Verification, Conclusion and Recommendation.

(1) Verification of the ranking of the significance of numerical data and drugs with research evidence

(a). Numerical features ranking verification:

For the rank of the features among the numerical data fields, it is reasonable that the number of emergency visits, outpatient visits and inpatient visits are the important features as compared to the rest since as mentioned in Statistical Brief (S. Claudia, B. Marguerite, and H. Katherine, May 2010), that many readmitted patients have preceding records of patient visits, either outpatient or inpatient. Further, the age factor should also be an important factor as it is one of the significant risk factors for the readmission (Miriam E. Tucker, July 05, 2013) and further, it is natural that the readmission probability of the older people will be different from that of the younger people.

(b). Bag of Drugs Ranking Verification:

We chose a research brochure developed by NaRCAD (the National Resource Center for Academic Detailing) with support from a grant from the Agency for Healthcare Research and Quality to the Division of Pharmacoepidemiology and Pharmacoeconomics of the Brigham and Women's Hospital Department of Medicine. It was adapted from materials originally developed by the non-profit Alosa Foundation.

This information is based on a comprehensive review of the evidence for best practices in the treatment of type 2 diabetes and is sponsored by the Agency for Health Care Research and Quality.

The key noteworthy points that play a part in introducing us to the treatment procedure/approach and verification of the nature of HBA1C results and recommendations are:

Table 1: Treatment Approach

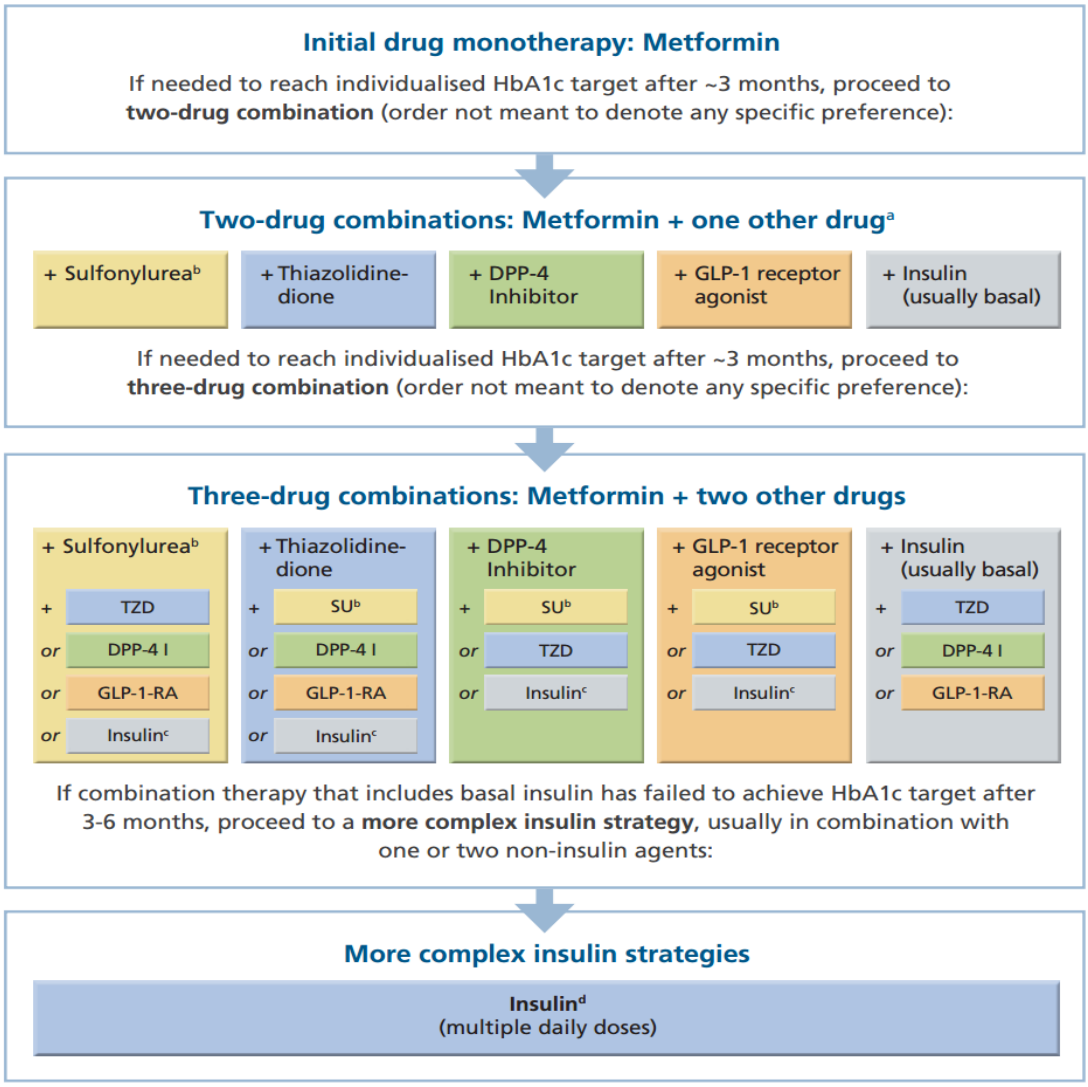
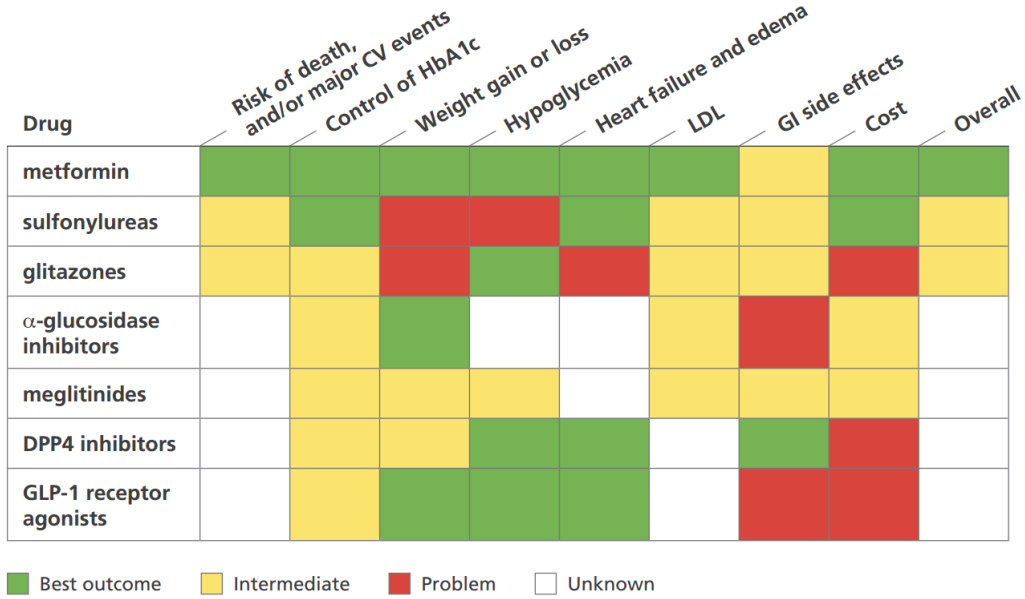


Table 2. Summary of comparative efficacy, safety, and cost of non-insulin agents



•If adequate control is not achieved after introducing a second line therapy, further intensification is necessary.

•Insulin offers the best chance to control HbA1c when added as a third agent, though non-insulin agents remain an option for patients unable or unwilling to use insulin.

-Clearly the two most important Drugs are the highest ranking drugs in our analysis as one of them :

Metformin: Is the first stage drug prescribed in initial drug monotherapy prescriptions and is a prime prescription.

and the other,

Insulin: Is the drug provided while practicing more complex insulin strategies and is extremely important if adequate control is not achieved after introducing a second line therapy in addition to initial drug monotherapy prescriptions.

The other high ranking drug names are either substitutions to the above drugs or prescriptions that are provided in combinations to these drugs as shown in the following table:

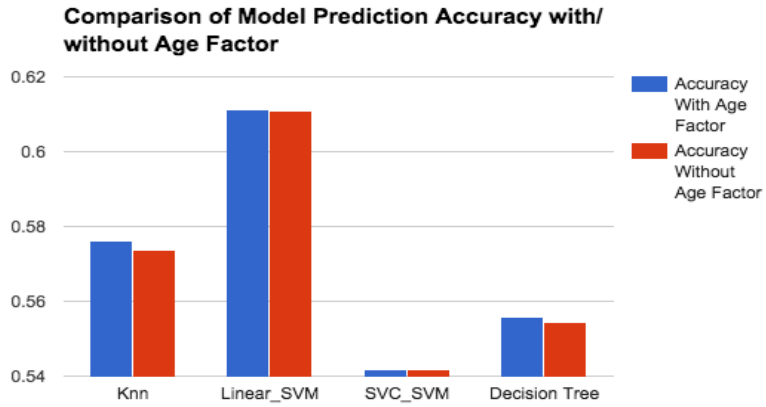
Important Index	Important Drugs (In decreasing importance)	Evidence Support Reference	Remarks
15	insulin	1,3,4,5	Insulin is taken only by injection. It's the majority of the diabetes drugs and type 1 diabetes patient must take insulin injection to compensate for the reduced ability of the pancreas.
0	metformin	2,3,4,5	Under Biguanides, widely used
1	repaglinide	3,4,5	Under Meglitinides
6	glipizide	2,3,4,5	Second generation under Sulfonylureas, widely used
4	glimepiride	2,3,4,5	Stronger Side effects than glipzide, second generation under Sulfonylureas
7	glyburide	3,4,5	Similar to glipzide, second generation under Sulfonylureas and widely used
11	acarbose	3,4	Under Alpha-glucosidase inhibitors
9	pioglitazone	3,4,5	(glitazones) Under Thiazolidinediones
10	rosiglitazone	3,4,5	(glitazones) Under Thiazolidinediones
12	miglitol	3,4	Under Alpha-glucosidase inhibitors
2	nateglinide	3,4,5	Under Meglitinides

Apparently, the top rank of importance shifts between Metmorfin and Insulin and is symbolizing that depending on the analysis or combination of data used its either the patients who need the requirement of initial drug monotherapy or those on whom on it doesn't work and are subjected to complex insulin strategies that dominate the data analytics section of drug combinations and is also verifying the hence obvious trend.

(2) Conslusion with respect to the scope

- Do specific drugs or combinations of drugs indicate likelihood of readmission?
  Yes, the dosage information on the bag of drags helps prediction of readmission rate. As illustrated in the various models built on the bag of drugs, the prediction accuracy can achieve up to 60% indicating that the information on the bag of drugs does help in predicting the readmission rate.

- How does age factor into the probability of readmission and/or outcomes?
  Since Age range turns out to be quite a significant feature as it is categorized in the second most important features set for the numerical data set; We conclude that age range does significantly factor into the prediction; Further, from the comparison of the models with and without age range as shown below, we could see that when age is not included, the accuracy in prediction decreases by 0.3% which is pretty significant as the prediction accuracy of the models are overall not very high. Thus, age factor is significant.



- Does the number of procedures, medication, and lab procedures correlate with either the readmission probability or the likelihood that the HbA1c test is performed?
  Yes, based on the observation that the prediction model built upon these numerical data features could go up to 61%, it is concluded that number of procedures, medications and lab procedures, etc., do correlate with the readmission probability.

- Do the categorical features of Admission Source, Admission Type and Discharge_disposition have a significant impact on the readmission prediction?
  Yes, as incorporating in the information of these categorical data in the prediction model of Decision tree significantly increases the prediction accuracy up to 1.2% thus, these categorical field of admission source, admission type and discharge_disposition do have an impact on the prediction accuracy and should be paid attention to.

(3) Recommendation for the hospital in readmission prediction for patients

- To predict the readmission rate and avoid extra cost due to excessed readmission rate of the patients, the hospital ought to carefully examine the historical data of patients and pay attention to certain fields.
- For instance, they should pay attention to fields like number of emergency visits, number of outpatient visits and number of emergency visits a patient had before the

year of encounter. Further, the age of the patient ought to be taken into consideration when predicting for readmission rate. In addition, hospital is also advised to consider the significance of admission_type , admission_source and discharge_disposition of the hospital and especially on discharge_disposition since it appears to be the most important categorical predictor among the three category fields. Thus, they are advised to not only pay attention to the inpatient care but also the continuing care after the patient is discharged.

- Lastly, certain drugs' dosage information should be carefully examined as they act as significant predictors for readmission rate. For instance, for drugs like metformin and insulin, dosage information should be carefully examined in predicting for readmission rate since they are significant predictors. For example, if one's dosage information on these drugs are significantly different from the other patients within the same group, then more attention should be paid to this patient.

6. Open problems/future research.

- Complex yet Better Preprocessing
  Better Runtime, Better Accuracy
- Incorporate HBA1C test result in the Prediction Model
  Better Accuracy
- Study the correlation between HBA1C test result and the readmission rate controlling the covariates of primary diagnosis result
  Verify the result generated in the research paper as mentioned in the section background information using a different approach.
- Integrate other Classification Models
  Based on other fields and do an Ensemble Prediction Model based on them
  Might make a difference for less frequent and more elaborate analysis
- Mortality Risk in Diabetes. Population-Based Case-Control Study : This study though, would require several other columns of data and information combined with the the information already present in the dataset.

7. Distribution of Work.

Yifan Xing (Research, Project Implementation, Presentation slides, Paper (Tools/Algorithms, Result & Analysis, Verification, Conclusion & Recommendation, Open problems/Future research))

Jai Sharma (Research, Presentation slides, Paper (Intro & Background, Data Description))

8. References

1. Dowshen, S. (2014, September 1). Medicines for Diabetes. Retrieved May 3, 2015, from http://kidshealth.org/teen/diabetes_center/treatment/medicines_diabetes.html#

2. Find the Best Medications for Type 2 Diabetes. (n.d.). Retrieved May 3, 2015, from http://www.healthline.com/health/consumer-reports-type-2-diabetes

3. What Are My Options? (n.d.). Retrieved May 3, 2015, from http://www.diabetes.org/living-with-diabetes/treatment-and-care/medication/oral-medications/what-are-my-options.html

4. Oral Diabetes Pills: Sulfonylureas, Biguanides, and More. (n.d.). Retrieved May 3, 2015, from http://www.webmd.com/diabetes/guide/oral-medicine-pills-treat-diabetes

5. Type 2 diabetes. (n.d.). Retrieved May 3, 2015, from http://www.mayoclinic.org/diseases-conditions/type-2-diabetes/basics/treatment/con-20031902

6. Improving Patient Care through Evidence: Treatment of Type 2 Diabetes (NaRCAD). Retrieved May 3, 2015, from http://www.narcad.org/wp-content/uploads/2013/02/NaRCAD_Diabetes_UnAd_final_2.25.pdf

7. Claudia Steiner, M.D., M.P.H., Marguerite Barrett, M.S., and Katherine Hunter, B.A. Statistical Brief #90. (n.d.). Retrieved May 3, 2015, from http://www.hcup-us.ahrq.gov/reports/statbriefs/sb90.jsp

8. Miriam E. Tucker, Medscape Log In. (n.d.). Retrieved May 3, 2015, from http://www.medscape.com/viewarticle/807384

9. Description of the Labels:
http://www.hindawi.com/journals/bmri/2014/781670/tab1/

10. Distribution of variable values and readmissions
http://www.hindawi.com/journals/bmri/2014/781670/tab3/

11. Category - code mapping of the primary diagnosis
http://www.hindawi.com/journals/bmri/2014/781670/tab2/

12. Source of data:
https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

13. Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records (Biomed Research International, 2014.). Retrieved May 4, 2015, from http://www.hindawi.com/journals/bmri/2014/781670/

9. Appendices

a. Index of bag of drugs

0 metformin
1 repaglinide
2 nateglinide
3 chlorpropamide
4 glimepiride
5 acetohexamide
6 glipizide
7 glyburide
8 tolbutamide
9 pioglitazone
10 rosiglitazone
11 acarbose
12 miglitol
13 troglitazone
14 tolazamide
15 insulin
16 glyburide-metformin
17 glipizide-metformin
18 glimepiride-pioglitazone
19 metformin-rosiglitazone
20 metformin-pioglitazone

b. Index of Numerical features

0 Age: ranges from the values (0-10) to (90-100), values can be normalized
1 Time spent in hospital: continuous but small range
2 Number of lab procedures: continuous values
3 Number of procedues: continuous values
4 Number of medications: continuous values
5 Number of Outpatient visits
6 Number of Inpatient visits
7 Number of Emergency visits