|  |
| --- |
| OSDS Optum-Scale Data Science 1909 Final Project |
| Diabetic Patient Readmission Data Analysis and Predication Modeling |
|  |

|  |
| --- |
| Yafen Huang  1/10/2019 |

**Table of contents**

1. Data source and lineage description
2. Context and background
3. Business problem and question
4. Value proposition
5. Hypothesis
6. Exploratory data analysis (EDA)
   1. Data quality and cleanup
   2. Data distribution
7. Setup for analysis
   1. Defining the variables and outcome
   2. Defining the baseline / control / denominator
   3. Data transformation / normalization as needed
   4. Sampling data as needed
   5. Rationale for method
8. Data analysis / modeling
   1. Core code segment for analysis / modeling (including method and parameters)
   2. Any intermediate results
   3. Table and plot of results
9. Summary of analysis result and explanation
10. Conclusion (accept or reject hypothesis)
11. Ethical considerations
12. Opportunity for improvement and further investigation
13. References
14. Appendix: Python notebook for modeling
    1. **Data Source and Lineage Description**
    2. **Dataset download site:**

The data csv file and description are provided on <https://github.optum.com/mkelly18/DSU-OSDS> (UHG internal access only) for the final project use. The original data and documentation can also be downloaded from UCI Machine Learning Repository <https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008>. The data are submitted on behalf of the Center for Clinical and Translational Research, Virginia Commonwealth University, a recipient of NIH CTSA grant UL1 TR00058 and a recipient of the CERNER data. John Clore(jclore@vcu.edu), Krzysztof J. Cios(kcios@vcu.edu), Jon DeShazo(jpdeshazo@vcu.edu), and Beata Strack([strackb@vcu.edu](mailto:strackb@vcu.edu)).

* 1. **Data collection and transformation:** 
     1. **Data collection**

The dataset was firstly created and used in the study “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records” [1]. It is an extract from the Health Facts database which contains data systematically collected from participating institutions electronic medical records. The database included 74,036,643 unique encounters (visits) that correspond to 17,880,231 unique patients and 2,889,571 providers when the study was conducted. Because it represents integrated delivery network health systems, the database contains both inpatient and outpatient data, including emergency department, for the same group of patients. The data extract used in the study represents 10 years (1999-2008) of clinical care at 130 US hospitals.

* + 1. **Initial dataset preparation**

To create the initial dataset that fits into the scope and boundary of the study, encounters with 55 attributes potentially associated with diabetic conditions or managements were extracted. These attributes describe patient demographics (gender, race and age), diagnoses, diabetic medications, number of visits in the year preceding the encounter, and payer information. Next, encounters are filtered by using the inclusion criteria listed as below:

* It is an inpatient encounter
* It is a diabetic encounter (diabetes is either a primary, secondary or third diagnosis)
* The length of stay was at least 1 day and at most 14 days
* Laboratory tests were performed during the encounter
* Medications were administered during the encounter

After applying filters, totally 101,766 encounters and 55 features were remained and saved as initial dataset. The dataset uploaded onto UCI Machine Learning Repository contains 101,766 encounters and 50 features. Based on the description on the UCI website, the dataset was filtered with same inclusion criteria as initial dataset in the study and thus can be considered at same stage of data lineage.

* 1. **Dataset cited in the previous studies**

When searching "diabetes 130 us hospitals for years 1999 2008" on Google Scholar [2], there are 73 articles cited this dataset. Some major usages can be categorized as below:

* Optimize diabetes patient readmission prediction with different classifiers and methodologies
* Developing new machine learning algorithm with improved classification accuracy.
* Developing classification model that analyzes the causes of HbA1c in diabetes patients
  1. **Context / background**

Total health expenditures have increased substantially in US over the past several decades [3]. One of the major factors that contribute to the rising cost is the increase in chronic disease such as diabetes. American Diabetes Association reported that patients with diabetes are 30 million representing about 9% of the population, but they account for approximately 25% of hospitalizations in US [4]. In 2012, costs associated with the hospitalization of diabetic patients were $124 billion and one of every five healthcare dollars is spent on patients with diabetes and its complications [5].

A major contributor of hospitalization high cost is the high number of readmission after discharge within 30 days. Readmission rate is a key measure of health care quality as well as a target for cost reduction. While the overall 30-day readmission rate of hospitalized patients is 8.5–13.5%, the 30-day readmission rate of diabetic patients is 14.4–22.7%. An estimated $25 billion of diabetic hospitalization cost was attributable to 30-day readmissions assuming a 20% readmission rate [6, 7]. Therefore reducing readmissions can greatly reduce healthcare costs as well as improve care for patients with diabetes. The Hospital Readmissions Reduction Program (HRRP) that penalizes hospitals with excessive readmission managed to decrease rates by 8 percent and save $2 billion to the Medicare program per between 2010 and 2015 [7]

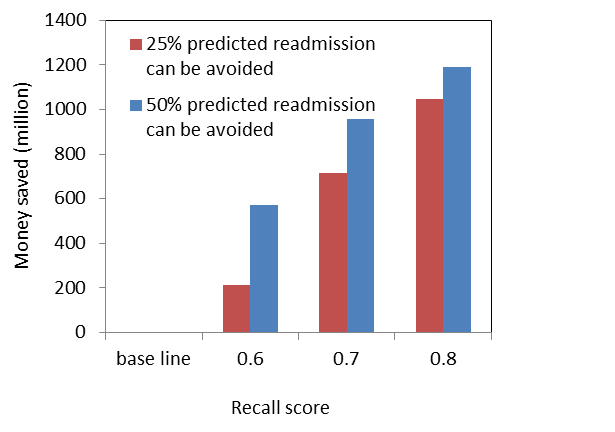
* 1. **Business problem and question**

In addition to penalty, early identification and targeting diabetic patients with high risk has also been proposed to have great potential in improving inpatient care and preventing readmission after discharge. Rational practices for identifying at-risk diabetic patients rely on clinicians’ assessment and decision of appropriate care plan for the individual patients. However these methods for determining readmission are only slightly better than random guessing [8].

Healthcare organizations are seeking opportunities to collect more actionable information from both patients and providers, and leverage the benefit of advanced data analytics to develop predictive intelligence so that patients with high risk of readmission can be captured with higher chance at early stage.

* 1. **Value proposition**

Based on the statistics from Healthcare Cost and Utilization Project (HCUP), in 2010 the number of Diabetes mellitus with complications readmission is 97,784 and readmission rate is 20.3% [8]. The average cost of readmission is $ 14,400. For this project, the targeted value is to correctly identify the patients with readmission risk and save money from avoidable readmission. Therefore the relationship of recall scores (TP/TP+FN) and money saved is proposed given a hypothetically 25%, 50% readmission avoidable rate as shown below.

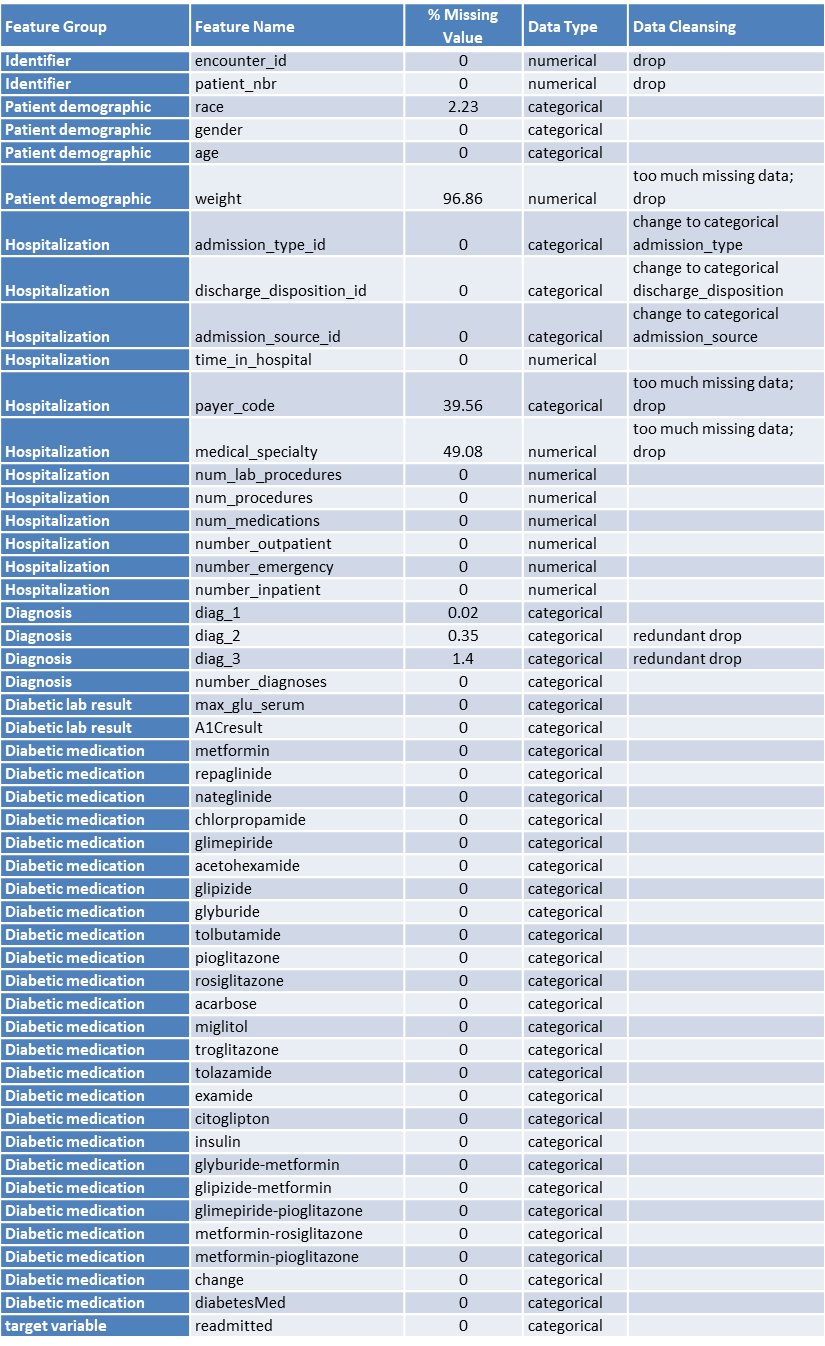


* 1. **Hypothesis**

In this project, with a diabetic patients readmission dataset that represents 10 years (1999-2008) of clinical care at 130 US and consists of attributes that describe patient demographics (gender, race and age), diagnoses, payer information, number of visits in the year preceding the encounter as well as diabetic specific medications [7], I am aiming to:

* 1. Identify key features that impact readmission of diabetic patients.
  2. Characterize whether diabetic related lab results and medication can help care providers optimize treatments and protocols, and make more informed decision before discharge and prevent readmission in the future.
  3. Create classifier model that can predict diabetic patients with high risk of readmission. Base model will be created with patient demographic, hospitalization and diagnosis feature to see how much readmission can be predicted by these features that are generic for all patients. Enhanced model will then be created with all the features used in base model, together with diabetic specific lab test and medication information to see whether adding diabetic specific features can improve model performance. Next, different classifiers including logistical regression, decision tree, random forest and SVM will be tested and evaluated to pick up one that yield highest accuracy and prediction power.
  4. **Exploratory data analysis (EDA)**
  5. **Data quality and cleanup**

The initial dataset downloaded from <https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008> contains 101,766 records with 50 features. The summary of feature format and percentage of missing values are listed below:



* + 1. **Deduplicate records.**

30,248 patients are found to have more than one encounter which cannot be considered as independent observations. Therefore only first encounters are kept for each patient to understand the primary admission and whether they are readmitted within 30 days.

* + 1. **Remove** **patients ended with hospice or death after discharge**

The encounters resulted in hospices or patient deaths don’t fit in the scope of this study on understanding readmission risk and thus are removed from dataset.

* + 1. **Replace features with numeric identifier with their full description**

Feature admission\_type, admission\_source, discharge\_disposition are coded in the numeric identifier admission\_type\_id, admission\_source\_id, discharge\_disposition\_id. New columns with mapped full descriptions are added and numeric identifiers are dropped afterwards.

* + 1. **Remove columns or rows with missing value**

Features ‘weight’, ’payer\_code’, ’medical\_specialty’ have great percentage of missing value and are dropped from dataset.

* + 1. **Remove outliers from numeric variable**

Outliers of numeric features 'time\_in\_hospital','num\_lab\_procedures', 'num\_medications' are removed based on IQR.

Summary of data cleansing steps and row/column count of each step:

|  |  |  |  |
| --- | --- | --- | --- |
| **Step** | **Process** | **Row count** | **Column count** |
| 1 | Initial dataset downloaded | 101766 | 50 |
| 2 | Remove duplicates of patients with multiple encounters | 71518 | 50 |
| 3 | Drop features weight, medical\_speciality,'payer\_code' with a lot of missing value | 71518 | 47 |
| 4 | Remove patients/encounters resulted in death | 70510 | 47 |
| 5 | Replace 'admission\_type\_id','discharge\_disposition\_id','admission\_source\_id' with full description | 70510 | 47 |
| 6 | Drop identifier column 'encounter\_id','patient\_nbr' | 70510 | 45 |
| 7 | Feature 'age' is binned to three buckets '<60','60-80','>80' | 70510 | 45 |
| 8 | Records with gender as 'unknown' are removed | 70507 | 45 |
| 9 | Remove outliers of 'time\_in\_hospital','num\_lab\_procedures','num\_medications' based on IQR | 66500 | 45 |
| 10 | Feature 'diag\_1' is binned to 8 buckets | 66500 | 45 |
| 11 | Feature 'A1Cresult' is binned to 'None','Norm','High' | 66500 | 45 |
| 12 | Feature'max\_glu\_serum' is binned to 'None','Norm','High' | 66500 | 45 |
| 13 | Add feature 'A1Ctaken' to indicate whether patient take A1C test or not | 66500 | 46 |
| 14 | Add feature 'dosage\_up', ‘dosage\_down','num\_dosage\_change','dosage\_change' to summarize the dosage change of all the diabetic medication | 66500 | 50 |
| 15 | Drop all the medications, diag\_2, diag\_3 | 66500 | 25 |

* 1. **Data distribution of feature variables and relationships with target variable**

Features included in this dataset represent information collected from patient, diagnosis, hospitalization, diabetic related lab result and medication. For better understanding how information from each aspect impact readmission rate, features are divided into four groups and data distribution and bivariate analysis are performed for features under each group:

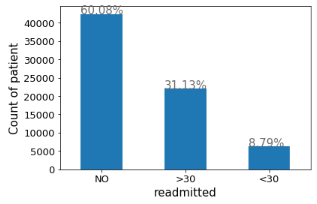
• Patient demographics

• Diagnosis

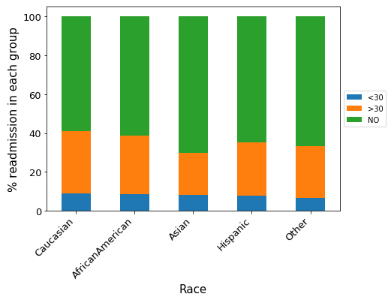
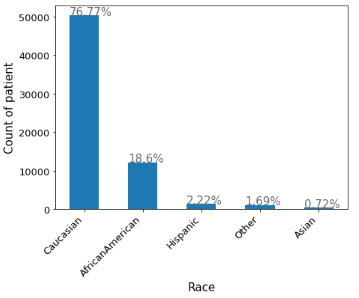
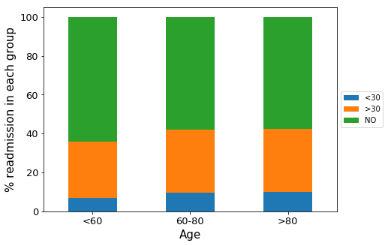
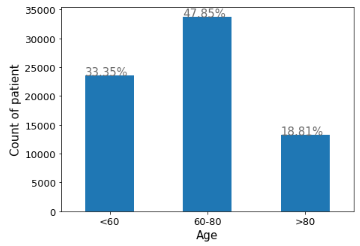
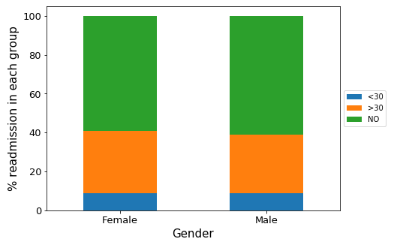
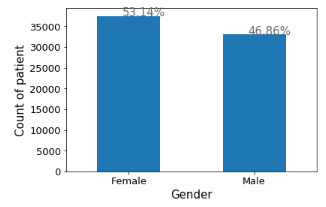
• Hospitalization

• Diabetic lab result and medication

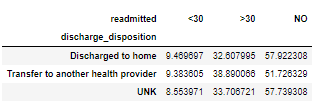
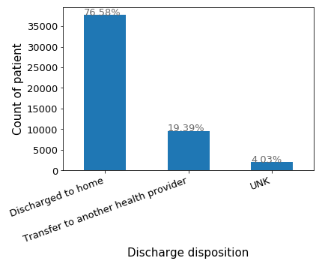
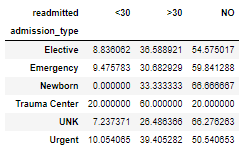
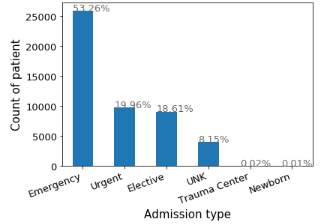
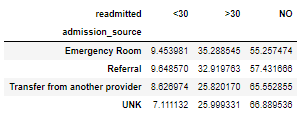
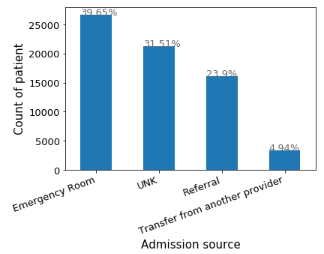
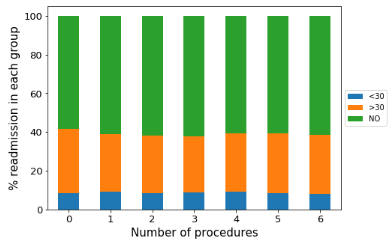
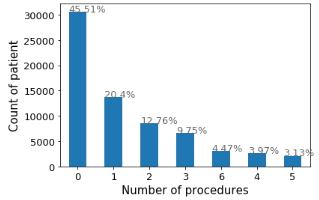
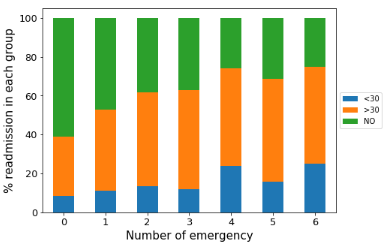
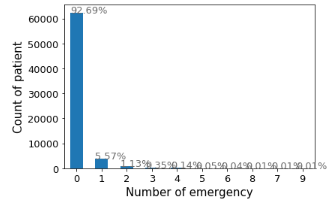
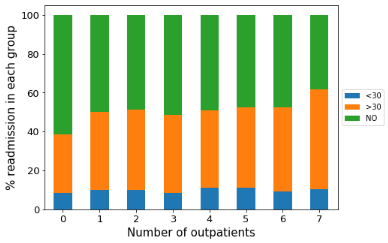
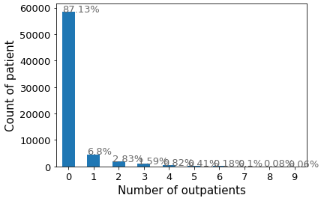
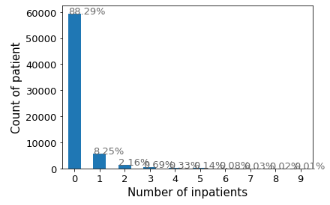
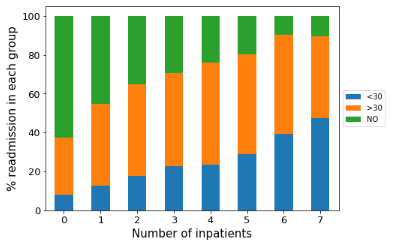
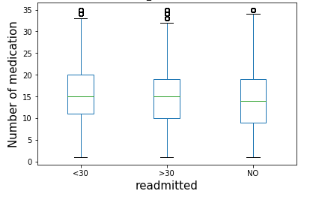
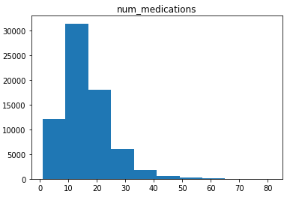
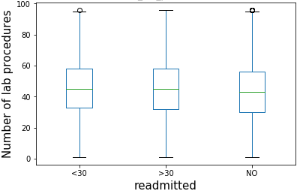
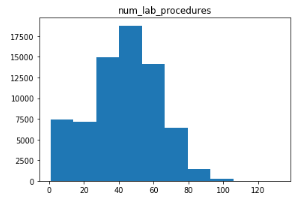
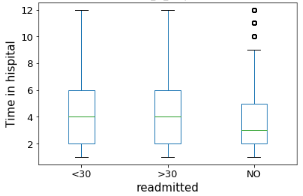
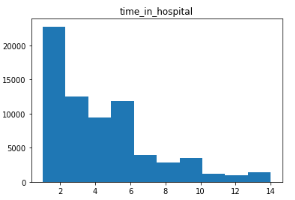
**6.2.1 Distribution of target variable ‘readmitted’**



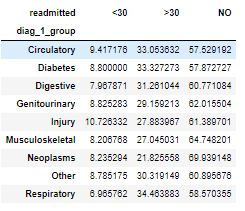
**6.2.2 Distribution of patient demographic features and relationship with readmission rate**

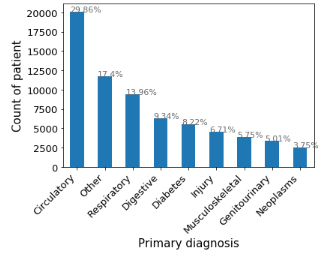


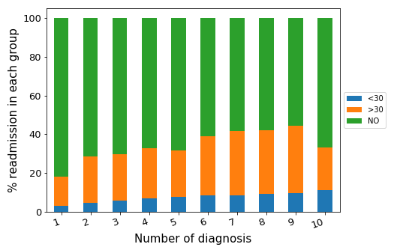
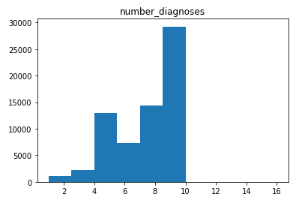
**6.2.3 Distribution of hospitalization features and relationship with readmission rate**



**6.2.4 Distribution of primary diagnosis and relationship with readmission rate**



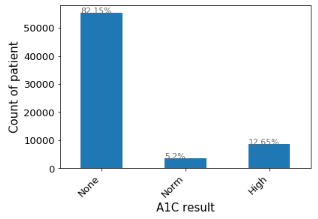




**6.2.5 Distribution of diabetic features and relationship with readmission rate**

**6.2.5.1 A1C test and readmission**

In this dataset, only 8.2% patients had diabetes as primary diagnosis and the rest had diabetes as secondary or third diagnosis. Around 17% A1C test was performed in the inpatient setting which cover patients with diabetes as primary or non-primary diagnosis.



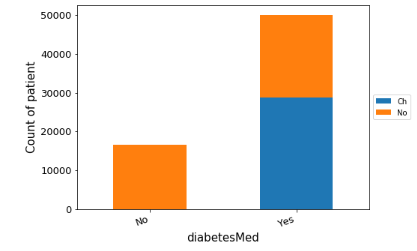
Previously, the study of impact of HbA1c measurement on hospital readmission rates by using the same dataset demonstrated that the relationship between probability of readmission and A1C result significantly depends on the primary diagnosis [1]. Readmission rates remained the highest for patients with circulatory diagnoses, but readmission rates for patients with diabetes appeared to be associated with the decision to test for A1C, rather than the values of the A1C result. Therefore, a new feature of whether A1C measure was taken or not is created and table below is the summary of the relationship of whether A1C test is taken and readmission rate in overall or broken down by disease type. The relationship of whether taking A1C test with the readmission risk also depends on the primary diagnoses. It significantly reduced readmission for patients with primary diagnosis as diabetes (15.31% vs. 9.73%), respiratory (11.31% vs. 7.87%) and injury (15.51% vs.10.28%), suggesting that increased attention and informed diabetic care can improve inpatient care for patients under these disease types. Circulatory is shown to have highest readmission rate but whether patients taking A1C test or not doesn’t affect readmission risk. One possible explanation is that hospitalization of chronic circulatory disease is very common and condition and care of primary circulatory disease weigh more in determining readmission risk than diabetes as non-primary disease. Overall, taking A1C test can lower readmission rate from 12.94% to 11.61%.

|  |  |  |  |
| --- | --- | --- | --- |
| **Primary diagnosis** | **A1Ctaken** | **Readmission within 30 days count** | **Readmission within 30 days rate %** |
| Circulatory | 0 | 1504 | 13.95 |
| 1 | 359 | 14.57 |
| Diabetes | 0 | 349 | 15.31 |
| 1 | 135 | 9.73 |
| Digestive | 0 | 438 | 11.74 |
| 1 | 58 | 10.6 |
| Genitourinary | 0 | 246 | 12.26 |
| 1 | 50 | 13.55 |
| Injury | 0 | 437 | 15.51 |
| 1 | 40 | 10.28 |
| Musculoskeletal | 0 | 275 | 11.17 |
| 1 | 33 | 11.91 |
| Neoplasms | 0 | 186 | 10.55 |
| 1 | 17 | 10.37 |
| Other | 0 | 847 | 12.78 |
| 1 | 177 | 11.83 |
| Respiratory | 0 | 554 | 11.31 |
| 1 | 95 | 7.87 |
| Overall | 0 | 1504 | 12.94 |
| 1 | 7304 | 11.61 |

**6.2.5.2 Diabetic medication change and readmission**

In this dataset, about 75% patients were treated with diabetic medication and among patients who took medication, 57.66% of them have medication type or dosage changed during hospitalization.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | | **Medication Change** | | **Total** |
| **Yes** | **No** |
| **Diabetes Medication** | **No** | **Count** | **0** | **16497** | **16497** |
| **Percent within diabetesMed=No** | 0% | 100% |  |
| **percent within total** |  | | 24.81% |
| **Yes** | **count** | 28835 | 21168 | 50003 |
| **Percent within diabetesMed=Yes** | 57.66% | 42.34% |  |
| **percent within total** |  | | 75.19% |
| **Total** | | | | | 66500 |



There are 23 features of different diabetic medications and dosage change in this dataset. These drugs all serve same purpose for controlling diabetes and the reason why different drugs were applied to different patients vary a lot across patients and beyond the scope of this study. Therefore we wouldn’t try to understand the impact of each drug individually on readmission and will only focus on number of medication dosage change that is relevant for this study. Patients who took diabetic medication might have more severity in diabetic condition and tend to have higher risk of readmission. Looking at the relationship between medication change and readmission, Change in medication increase readmission risk in almost all primary diagnosis groups in a quantitatively dependent manner, suggesting that if it is not targeted treatment for primary diabetes, maintaining same treatment is important for stabilizing diabetes and minimize the side effect on primary disease treatment and patients’ health condition after discharge.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Primary diagnosis** | **Diabetes Medication** | **Medication change** | **number of medication dosage change** | **Readmission within 30 days count** | **Readmission within 30 days percent%** |
| Circulatory | Yes | Ch | 3 | 2 | 20 |
| 2 | 21 | 16.03 |
| 1 | 439 | 17.83 |
| 0 | 379 | 13.8 |
| No | 0 | 642 | 14.77 |
| No | No | 0 | 380 | 10.71 |
| Diabetes | Yes | Ch | 4 | 0 | 0 |
| 3 | 3 | 23.08 |
| 2 | 10 | 13.16 |
| 1 | 187 | 12.41 |
| 0 | 70 | 12.8 |
| No | 0 | 136 | 13.13 |
| Digestive | No | No | 0 | 78 | 16.05 |
| Yes | Ch | 3 | 0 | 0 |
| 2 | 4 | 25 |
| 1 | 89 | 13.24 |
| 0 | 70 | 11.86 |
| No | 0 | 187 | 13.06 |
| No | No | 0 | 146 | 9.31 |
| Genitourinary | Yes | Ch | 3 | 0 | 0 |
| 2 | 7 | 24.14 |
| 1 | 84 | 17.46 |
| 0 | 45 | 10.25 |
| No | 0 | 102 | 13.25 |
| No | No | 0 | 58 | 8.84 |
| Injury | Yes | Ch | 3 | 1 | 100 |
| 2 | 3 | 11.54 |
| 1 | 118 | 17.48 |
| 0 | 109 | 15.64 |
| No | 0 | 157 | 16.42 |
| No | No | 0 | 89 | 10.45 |
| Musculoskeletal | Yes | Ch | 3 | 2 | 66.67 |
| 2 | 4 | 14.81 |
| 1 | 54 | 11.82 |
| 0 | 89 | 11.19 |
| No | 0 | 103 | 12.22 |
| No | No | 0 | 56 | 9.14 |
| Neoplasms | Yes | Ch | 2 | 4 | 21.05 |
| 1 | 54 | 15.25 |
| 0 | 35 | 9.67 |
| No | 0 | 65 | 11.04 |
| No | No | 0 | 45 | 7.46 |
| Other | Yes | Ch | 4 | 0 | 0 |
| Ch | 3 | 1 | 16.67 |
| 2 | 20 | 22.73 |
| 1 | 242 | 14.28 |
| 0 | 195 | 12.12 |
| No | 0 | 327 | 13.02 |
| No | No | 0 | 239 | 10.81 |
| Respiratory | Yes | Ch | 3 | 0 | 0 |
| 2 | 4 | 5.13 |
| 1 | 168 | 12.71 |
| 0 | 139 | 10.72 |
| No | 0 | 219 | 11.56 |
| No | No | 0 | 119 | 7.89 |

In summary whether patients get readmission is a consequence of patient’s health condition, disease severity, complexity and quality of inpatient care in combined. Taking A1C test and change in the diabetic medication have significant impact on the readmission risk, but the influence seems to be context dependent and there is no single pattern that fit all the disease types. When diabetes plays major role in patients’ health condition during hospitalization and after discharge, diabetic related measurement would help make more informed decision for health care providers and reduce readmission. Overall the pattern of how diabetic related features impact readmission probability provides useful information for improving care quality and preventing avoidable readmission after discharge. Given the complexity of disease condition and clinical practice, it would be better if the contribution diabetic related information could be evaluated under the setting with more medical and clinical knowledge.

1. **Setup for analysis**
   1. **Defining variables and outcome**

For developing model that can predict the probability of readmission for diabetic patients, the target variables would be ‘readmitted’. Features that represent patients demographic, diagnosis, hospitalization, lab result and medication would be combined together to train the model.

* 1. **Defining the baseline / control / denominator**

Target variable ‘readmitted’ has three categories readmission within 30 days, readmission greater than 30 days and no readmission. By looking at the pattern of the relationship between different features with readmission, the readmission greater than 30 days category is very often intermediate state between readmission within 30 days and no readmission, and thus confuses the boundary between the two. Meanwhile, opinions from medical experts also suggest that readmission greater than 30 days is very likely due to conditions other than first encounter at hospital and would not contribute to the understanding and prediction of patients of readmission within 30 days and those who never get readmitted. Therefore readmission greater than 30 days is considered to be removed for future analysis and modeling. The binary target variable was then balanced by using undersampling method as oversampling compromise independence between training and test set.

From ethical perspective, potential bias from patient demographic features such as gender and race etc. also needed to be considered before feeding into modeling. Feature ‘gender’ was balanced with equal number of female and male patients to avoid bias on the true population. I originally planned to balance feature ‘race’ and ‘age’ by using undersampling as well. However, these two features are very unbalanced and balancing using undersampling will greatly compromise the size of dataset and statistical power. Doing multivariate analysis of this age and race balanced dataset also showed that pattern change compared with original dataset. Therefore, I decided to drop race and age feature instead of keeping and balancing it.

|  |  |  |
| --- | --- | --- |
| **Model** | **Feature collection** | **Denominator** |
| **Base model** | Non-diabetic specific features   * patient demographic * hospitalization * primary diagnosis | 1.     Remove readmitted>30;  2.    Balance readmitted <30 and NO  3.    Balance gender  4. Remove race and age |
| **Enhanced model** | Non-diabetic specific features   * patient demographic * hospitalization * primary diagnosis   Diabetic related features   * Lab result * Medication |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

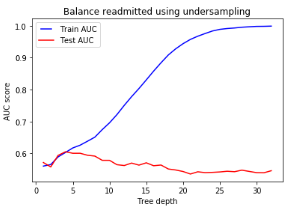
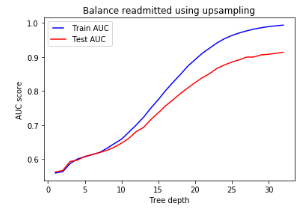
* 1. **Data transformation and normalization**

Below are the steps of transformation and normalization for preparing data for modeling.

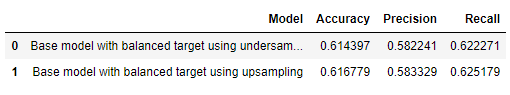
|  |  |  |
| --- | --- | --- |
| **Steps** | **Feature** | **Transformation** |
| 1 | readmitted | '<30' =1;'No'=0 |
| 2 | change | ‘Ch'=1;'No'=0 |
| 3 | diabeteMed | ‘Yes'=1;'No'=0 |
| 4 | gender, discharge\_disposition, admission\_type, admission\_source, diag\_1\_group | transform to dummy variables |
| 5 | Normalize all the features | scaler.fit\_transform(training set) |

* 1. **Sampling data for balancing features**

For balancing the target variable ‘readmitted’, both undersampling and upsampling were tested and compared with logistic regression and decision tree method. By using decision tree method, the plot of tree depth and AUC score for the train and test set prepared with upsampling method show similar AUC score trend as tree depth increases, suggesting that the test and train test are very similar and not independent with each other. The trending of AUC score for train and test set prepared with undersampling is more independent and make sense when tree depth increases.

Upsampling seems not increase performance in logistic regression as well as shown below. Therefore undersampling method is decided to be used for balancing feature for this study.



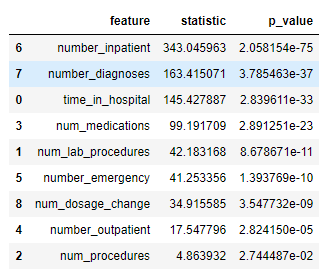
* 1. **Rationale for method**

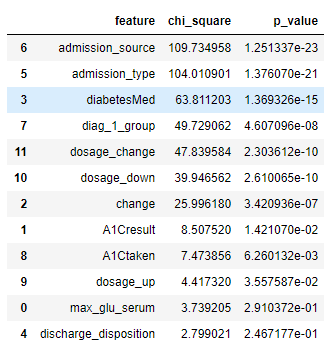
Algorithm of logistic regression is relatively simple to understand the contribution of each feature and usually works well when independent variables are a mixture of numeric and nominal data. Therefore for understanding the contribution of each base feature to and how adding diabetic related features influence model performance, logistic regression was used to perform feature curation and comparison between base model and enhanced model. For the final model, classifiers such as decision tree, random forest and SVM were also tested and compared for the model performance. Tenfold (k=10) cross validation was used for evaluating the models to reduce the bias from training set.

1. **Data analysis / modeling**
   1. **Core code segment for analysis / modeling**

The python notebook that contains core code for analysis and modeling is attached at the end of the report.

* 1. **Results of modeling** 
     1. **Statistical analysis of features and target variable**

Anova one way test of numeric features with target variable ‘readmitted’

Chi square test of numeric features with target variable ‘readmitted’

* + 1. **Development of base model by logistic regression**

Firstly, base model with all the variables except diabetic related features were fed into logistic regression. The accuracy, precision, and recall for this model are 61.31%, 57%, 62.4% respectively. Dropping one feature at a time does not improve the performance of base model significantly. Therefore all the base features are kept for the final base model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Model** | **Accuracy** | **Precision** | **Recall** |
| Logistic | Base model with all base features | 0.6131 | 0.57 | 0.624 |
| Logistic | Base model with dropping gender | 0.6135 | 0.5695 | 0.6246 |
| Logistic | Base model with dropping time\_in\_hospital | 0.6098 | 0.5678 | 0.6198 |
| Logistic | Base model with dropping num\_lab\_procedures | 0.6145 | 0.5717 | 0.6255 |
| Logistic | Base model with dropping num\_procedures | 0.6098 | 0.5665 | 0.6203 |
| Logistic | Base model with dropping num\_medications | 0.6134 | 0.5708 | 0.6244 |
| Logistic | Base model with dropping number\_outpatient | 0.6133 | 0.5708 | 0.6241 |
| Logistic | Base model with dropping number\_emergency | 0.611 | 0.5678 | 0.6217 |
| Logistic | Base model with dropping number\_inpatient | 0.5971 | 0.5942 | 0.598 |
| Logistic | Base model with dropping number\_diagnoses | 0.6034 | 0.538 | 0.619 |
| Logistic | Base model with dropping discharge\_disposition | 0.6112 | 0.5631 | 0.6232 |
| Logistic | Base model with dropping admission\_type | 0.6124 | 0.5704 | 0.623 |
| Logistic | Base model with dropping admission\_source | 0.6138 | 0.5708 | 0.6247 |
| Logistic | Base model with dropping diag\_1\_group | 0.6084 | 0.5683 | 0.6179 |

Due to the heterogeneity and complexity of different disease causes and hospitalization conditions, the model performance is also evaluated per each primary diagnosis. Model accuracy ranges from 66.31 to 55.26%, which is relatively higher for patients with primary diagnosis as diabetes, digestive and lowest for patients with injury and neoplasms, suggesting there is heterogeneity of model performance among different primary diagnosis as well.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classifier** | **Model** | **Primary diagnosis** | **Accuracy** | **Precision** | **Recall** |
| Logistic | Base model | Diabetes | 0.6631 | 0.6322 | 0.6801 |
| Logistic | Base model | Digestive | 0.6376 | 0.4345 | 0.6425 |
| Logistic | Base model | Respiratory | 0.6279 | 0.5267 | 0.6425 |
| Logistic | Base model | Musculoskeletal | 0.6145 | 0.4468 | 0.62 |
| Logistic | Base model | Other | 0.6131 | 0.5238 | 0.6291 |
| Logistic | Base model | Genitourinary | 0.6103 | 0.5796 | 0.6046 |
| Logistic | Base model | Circulatory | 0.6028 | 0.6684 | 0.6216 |
| Logistic | Base model | Injury | 0.5756 | 0.6665 | 0.5943 |
| Logistic | Base model | Neoplasms | 0.5526 | 0.2486 | 0.4872 |

* + 1. **Development of enhanced model with adding diabetic related features**

To evaluate the impact of adding diabetic related features on predicting readmission, diabetic related features were added to modeling sequentially and mode performance is compared and evaluated between the models. Adding diabetic related features doesn’t improve the model performance significantly in overall. I also tried to denominate dataset by removing the A1Cresult as ‘None’, compared enhanced model with diabetic features vs. based model for only patients with A1C results and didn’t see significant improvement as well (data not shown here).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Model** | **Accuracy** | **Precision** | **Recall** |
| Logistic | Base model | 0.6131 | 0.57 | 0.624 |
| Logistic | Enhanced model with adding A1Cresult | 0.6139 | 0.573 | 0.6241 |
| Logistic | Enhanced model with adding A1Ctaken | 0.6136 | 0.5732 | 0.6237 |
| Logistic | Enhanced model with A1Ctaken and diabetesMed | 0.6147 | 0.5762 | 0.6245 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed,change | 0.6139 | 0.5758 | 0.6235 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed,change, num\_dosage\_change (final enhanced model) | 0.6134 | 0.5726 | 0.6237 |

If looking at model performance through each primary diagnosis, the impact of adding those diabetic feature seems to be positive (in green) on model performance in some disease type such as diabetic, neoplasm and negative (in red) in others such as digestive and musculoskeletal. It should be noted that the percentage of patients who had A1C test taken, or had multiple change on medication dosage is very low in general. Even though we have seen relationship between these features with readmission risk, how much capturing these readmitted patients correctly can contribute to model accuracy still depends on the percentage of those patients in total population.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classifier** | **Model** | **Primary diagnosis** | **Accuracy** | **Precision** | **Recall** |
| Logistic | Base model | Circulatory | 0.6028 | 0.6684 | 0.6216 |
| Logistic | Enhanced model with A1Cresult | Circulatory | 0.6013 | 0.6678 | 0.62 |
| Logistic | Enhanced model with A1Ctaken | Circulatory | 0.6009 | 0.6665 | 0.6199 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed | Circulatory | 0.6046 | 0.6758 | 0.6218 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed,change | Circulatory | 0.6043 | 0.6752 | 0.6214 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed,change,num\_dosage\_change | Circulatory | 0.6064 | 0.6766 | 0.6234 |
| Logistic | Base model | Diabetes | 0.6631 | 0.6322 | 0.6801 |
| Logistic | Enhanced model with A1Cresult | Diabetes | 0.6719 | 0.6428 | 0.6876 |
| Logistic | Enhanced model with A1Ctaken | Diabetes | 0.6731 | 0.6449 | 0.6889 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed | Diabetes | 0.6612 | 0.6157 | 0.6848 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed,change | Diabetes | 0.6554 | 0.6084 | 0.6793 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed,change,num\_dosage\_change | Diabetes | 0.6571 | 0.6138 | 0.6801 |
| Logistic | Base model | Digestive | 0.6279 | 0.5267 | 0.6425 |
| Logistic | Enhanced model with A1Cresult | Digestive | 0.6227 | 0.5215 | 0.6331 |
| Logistic | Enhanced model with A1Ctaken | Digestive | 0.624 | 0.5235 | 0.6338 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed | Digestive | 0.6177 | 0.5267 | 0.627 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed,change | Digestive | 0.6177 | 0.5291 | 0.627 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed,change,num\_dosage\_change | Digestive | 0.6134 | 0.5202 | 0.6229 |
| Logistic | Base model | Genitourinary | 0.6103 | 0.5796 | 0.6046 |
| Logistic | Enhanced model with A1Cresult | Genitourinary | 0.6165 | 0.5854 | 0.6114 |
| Logistic | Enhanced model with A1Ctaken | Genitourinary | 0.6137 | 0.5767 | 0.6109 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed | Genitourinary | 0.6117 | 0.5688 | 0.6045 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed,change | Genitourinary | 0.6171 | 0.576 | 0.61 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed,change,num\_dosage\_change | Genitourinary | 0.6049 | 0.5664 | 0.5976 |
| Logistic | Base model | Injury | 0.5756 | 0.6665 | 0.5943 |
| Logistic | Enhanced model with A1Cresult | Injury | 0.5781 | 0.671 | 0.5955 |
| Logistic | Enhanced model with A1Ctaken | Injury | 0.5768 | 0.671 | 0.5944 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed | Injury | 0.5942 | 0.6681 | 0.6164 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed,change | Injury | 0.5942 | 0.6673 | 0.6157 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed,change,num\_dosage\_change | Injury | 0.5916 | 0.6534 | 0.6157 |
| Logistic | Base model | Musculoskeletal | 0.6145 | 0.4468 | 0.62 |
| Logistic | Enhanced model with A1Cresult | Musculoskeletal | 0.6167 | 0.4461 | 0.6216 |
| Logistic | Enhanced model with A1Ctaken | Musculoskeletal | 0.6125 | 0.4499 | 0.6129 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed | Musculoskeletal | 0.6084 | 0.4359 | 0.6088 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed,change | Musculoskeletal | 0.6073 | 0.4328 | 0.6075 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed,change,num\_dosage\_change | Musculoskeletal | 0.6066 | 0.4321 | 0.6045 |
| Logistic | Base model | Neoplasms | 0.5526 | 0.2486 | 0.4872 |
| Logistic | Enhanced model with A1Cresult | Neoplasms | 0.5457 | 0.2395 | 0.4683 |
| Logistic | Enhanced model with A1Ctaken | Neoplasms | 0.5502 | 0.25 | 0.4812 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed | Neoplasms | 0.5729 | 0.3144 | 0.5257 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed,change | Neoplasms | 0.566 | 0.3085 | 0.5099 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed,change,num\_dosage\_change | Neoplasms | 0.5721 | 0.3026 | 0.5185 |
| Logistic | Base model | Other | 0.6131 | 0.5238 | 0.6291 |
| Logistic | Enhanced model with A1Cresult | Other | 0.6145 | 0.5329 | 0.6284 |
| Logistic | Enhanced model with A1Ctaken | Other | 0.6145 | 0.5343 | 0.6279 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed | Other | 0.6123 | 0.5356 | 0.6262 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed,change | Other | 0.6113 | 0.5335 | 0.6248 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed,change,num\_dosage\_change | Other | 0.6118 | 0.5306 | 0.626 |
| Logistic | Base model | Respiratory | 0.6376 | 0.4345 | 0.6425 |
| Logistic | Enhanced model with A1Cresult | Respiratory | 0.6386 | 0.4388 | 0.6413 |
| Logistic | Enhanced model with A1Ctaken | Respiratory | 0.6379 | 0.4371 | 0.6413 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed | Respiratory | 0.6386 | 0.4476 | 0.6386 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed,change | Respiratory | 0.6381 | 0.4524 | 0.6362 |
| Logistic | Enhanced model with A1Ctaken,diabetesMed,change,num\_dosage\_change | Respiratory | 0.6365 | 0.4462 | 0.6354 |

* + 1. **Enhanced model with other classifiers**

In addition to logistic regression, other classifiers such as decision tree, random forest, and SVM have also been tested for the enhanced model as shown below. Parameters for each classifier were optimized but no big difference of performance was found between them.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Model** | **Accuracy** | **Precision** | **Recall** |
| Logistic | Enhanced model with diabetic features | 0.6134 | 0.5726 | 0.6237 |
| Decision tree | Enhanced model with diabetic features | 0.6044 | 0.5832 | 0.6092 |
| Random forest | Enhanced model with diabetic features | 0.6131 | 0.5723 | 0.6234 |
| SVM | Enhanced model with diabetic features | 0.6136 | 0.5479 | 0.6309 |

**9. Summary of analysis result and explanation**

Results from data exploratory analysis and prediction modeling suggest that in this diabetic readmission dataset, majority of the features from patient demographic, hospitalization, diagnosis, diabetic test result and medication can provide relevant information on predicting readmission. Getting readmitted is a consequence of patient’s health condition, disease severity and complexity, and quality of care from providers in combined. Diabetic related measurement and medication info can provide important guidance for improving inpatient care and prevent avoidable readmission in a primary diagnosis dependent manner.

Base model with patient demographic, hospitalization and diagnosis feature can give 61% accuracy for predicting readmission within 30 days. Model scores partitioned by primary diagnosis suggest heterogeneity of performance among different disease which is not surprising given the heterogeneity of disease causes and conditions. Enhanced model with adding diabetic related features doesn’t improved model performance in overall, but is found to improve prediction in some primary disease such as diabetes, injury and cancer. Comparing with readmission risk from the patient health condition and disease severity, the practices of lab testing and medication are relatively easier to implement, improve and bring impact for reducing readmission. However it should be noted that the extent that improving diabetic related care can reduce readmission risk also depends on primary diagnosis and how much diabetes weigh in patients’ condition during hospitalization.

**10. Conclusion (accepts or rejects hypothesis)**

Based on the result from data analysis and modeling, the conclusions for proposed hypothesis are:

* 1. Identify key features that impact readmission of diabetic patients.

*Conclusion: accepted. A list of key features from hospitalization, diagnosis and diabetic lab result and medication are found to influence the probability of readmission.*

* 1. Check whether diabetic specific lab results and medication can help care providers optimize treatments and protocols, and make more informed decision before discharge and prevent readmission in the future.

*Conclusion: accepted. Taking A1C test and avoiding non-targeted/unnecessary medication change are found to lower the readmission risk and can be used as guidance for improving inpatient care.*

* 1. Create classifier model that can predict diabetic patients with high risk of readmission. Base model will be created with patient demographic, hospitalization and diagnosis feature to see how much readmission can be predicted by these features that are generic for patients. Enhanced model will then be created with all the features used in base model, together with diabetic specific lab test and medication information to see whether adding diabetic specific features can improve model performance.

*Conclusion: accepted. Results from multiple classifiers are all shown to have reasonable prediction accuracy, precision better than rational practices that rely on clinicians’ assessment and decision of appropriate care plan for identifying at-risk diabetic patients [8]*

**11. Ethical considerations**

In order to avoid potential bias, patients’ demographic features gender, race and age were either balanced or removed from prediction modeling.

**12. Opportunity for improvement and further investigation**

1. Work on dataset with more diabetic related test so that the analysis could be more informed.

2. Optimize the model prediction performance.

3. Work on dataset that is more updated and reflected current trending in real world.

## 13. Reference

1. Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records,” BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.
2. Google scholar search <https://scholar.google.com/scholar?start=60&q=%22diabetes+130+us+hospitals+for+years+1999+2008%22&hl=en&as_sdt=0,34>
3. <https://www.thebalance.com/causes-of-rising-healthcare-costs-4064878>
4. http://www.diabetes.org /diabetes basics/statistics/ infographics adv staggering cost of diabetes.html
5. American Diabetes Association (2013). “Economic Costs of Diabetes in the U.S. in 2012”
6. Current Diabetes Reports (2018) “Correction to: Hospital Readmission of Patients with Diabetes”
7. http://www.medpac.gov/-blog-/the-hospital-readmissions-reduction-program-(hrrp)-has-succeeded-for-beneficiaries-and-the-medicare-program/2018/06/15/the-hospital-readmissions-reduction-program-has-succeeded-for-beneficiaries-and-the-medicare-program
8. Allaudeen, Nazima, Jeffrey L. Schnipper, E. John Orav, Robert M. Wachter, and Arpana R. Vidyarthi. (2011). “Inability of Providers to Predict Unplanned Readmissions.” Journal of General Internal Medicine 26(7):771–76.
9. <https://www.hcup-us.ahrq.gov/reports/statbriefs/sb196-Readmissions-Trends-High-Volume-Conditions.pdf>