Institute of Data

PREDICTIVE MODEL FOR HOSPITAL READMISSIONS

Jane Tran

Capstone Project
April 2023

PROJECT CONTEXT

Industry or domain

Reducing diabetic readmissions can reduce medical cost and improve patients' wellbeing as diabetes is a leading cause of death and costly chronic disease in the U.S.

Problem area

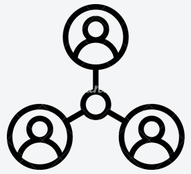
To identify patients with high probability of readmission based on initial diagnoses, number of procedures, and other variables

Previous work in this area

- HbA1c measurement analysis improves patient outcomes and minimizes inpatient care costs for diabetes-related hospital readmissions (Strack et al., 2014).
- Risk factors for hospital readmission in diabetic patients include lower socioeconomic level, racial minority status, burden, insurance status, emergent admission, and recent hospitalization history (Rubin, 2015).

DEFINE

BUSINESS UNDERSTANDING



Stakeholders

- Healthcare providers (physicians, nurses, doctors, etc.)
- Hospital Administrators
- Patients & Family
- Payers & Insurers
- Researchers and academics
- Policy makers and regulators
- Technology vendors and developers

Business Question

- How to identify patients with high probability of readmission based on initial diagnoses, number of procedures, and other variables?

Business Value

- The projected cost savings of this project for the healthcare sector are estimated to be $27 billion *

* See slide 16 for detailed calculation

DATA UNDERSTANDING

Data Questions

- Does diabetes play a central role in readmission?
- On what groups of patients should the hospital focus their follow-up efforts to better monitor patients with a high probability of readmission?

Dataset Description

- Source: Diabetes 130-US hospitals for years 1999-2008 Data Set
- Contains 101,766 observations and 50 features
- Target Feature: 'Readmitted' within 30 days

Feature name	Type	% missing	Description and values
Encounter ID	Numeric	0%	Unique identifier of an encounter
Patient number	Numeric	0%	Unique identifier of a patient
Race	Nominal	2%	Values: Caucasian, Asian, African American, Hispanic, and other
Gender	Nominal	0%	Values: male, female, and unknown/invalid
Age	Nominal	0%	Grouped in 10-year intervals: 0, 10), 10, 20), ..., 90, 100)
Weight	Numeric	97%	Weight in pounds.
Admission type	Numeric	0%	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available
Discharge disposition	Numeric	0%	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
Admission source	Numeric	0%	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital
Time in hospital	Numeric	0%	Integer number of days between admission and discharge
Payer code	Nominal	52%	Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay
Medical specialty	Nominal	53%	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon
Number of lab procedures	Numeric	0%	Number of lab tests performed during the encounter
Number of procedures	Numeric	0%	Number of procedures (other than lab tests) performed during the encounter
Number of medications	Numeric	0%	Number of distinct generic names administered during the encounter
Number of outpatient visits	Numeric	0%	Number of outpatient visits of the patient in the year preceding the encounter
Number of emergency visits	Numeric	0%	Number of emergency visits of the patient in the year preceding the encounter
Number of inpatient visits	Numeric	0%	Number of inpatient visits of the patient in the year preceding the encounter
Diagnosis 1	Nominal	0%	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values
Diagnosis 2	Nominal	0%	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values
Diagnosis 3	Nominal	1%	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values
Number of diagnoses	Numeric	0%	Number of diagnoses entered to the system
Glucose serum test result	Nominal	0%	Indicates the range of the result or if the test was not taken. Values: ">200," ">300," "normal," and "none" if not measured
A1c test result	Nominal	0%	Indicates the range of the result or if the test was not taken. Values: ">8" if the result was greater than 8%, ">7" if the result was greater than 7% but less than 8%, "normal" if the result was less than 7%, and "none" if not measured.
Change of medications	Nominal	0%	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: "change" and "no change"
Diabetes medications	Nominal	0%	Indicates if there was any diabetic medication prescribed. Values: "yes" and "no"
23 features for medications	Nominal	0%	For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: "up" if the dosage was increased during the encounter, "down" if the dosage was decreased, "steady" if the dosage did not change, and "no" if the drug was not prescribed
Readmitted	Nominal	0%	Days to inpatient readmission. Values: "<30" if the patient was readmitted in less than 30 days, ">30" if the patient was readmitted in more than 30 days, and "No" for no record of readmission.

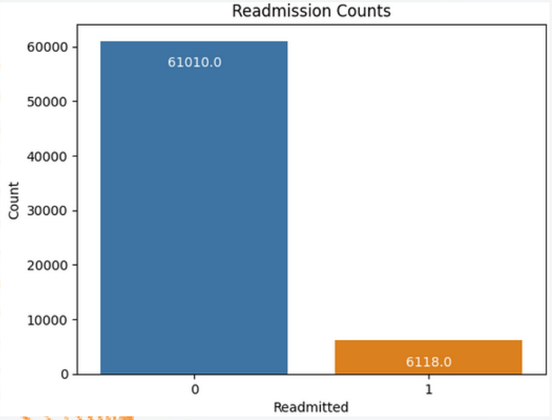
DESIGN

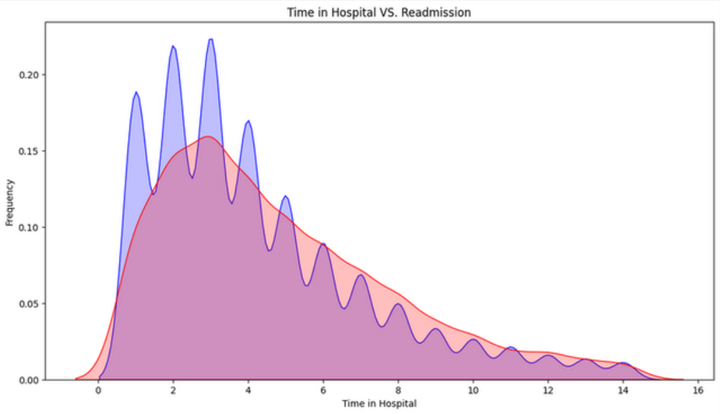
Institute of Data

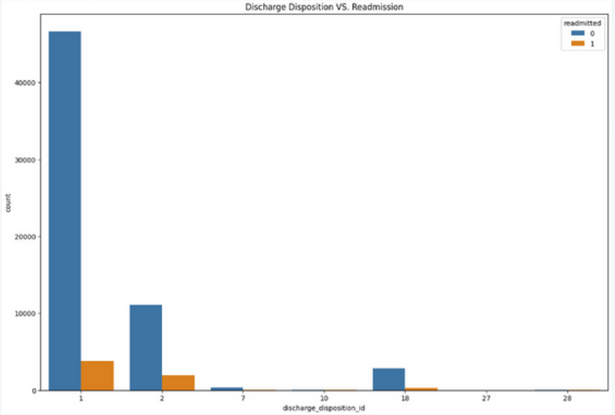
DATA PREPROCESSING

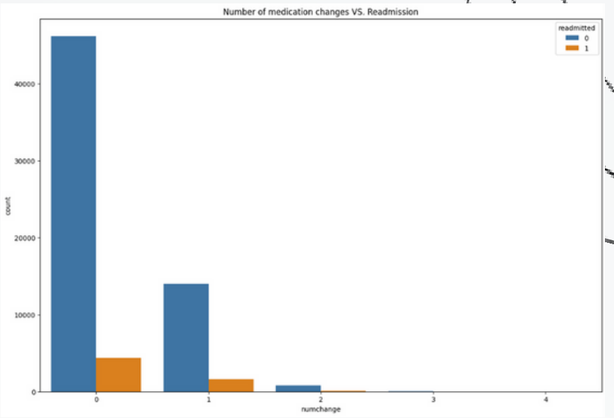
- Missing values were either dropped or encoded as Unknown
- Complexed features were recoded and simplified such as admission types, discharge disposition, diagnoses
- Categorical features were encoded using one-hot encoding technique
- After train test spliting, standardization and outlier detection methods were used
- Since this is a imbalanced dataset (readmitted only accounted 11%), a resampling technique called Synthetic Minority Oversampling Technique + Edited Nearest Neighbors were applied so as the models perform to their best and give most accurate predictions

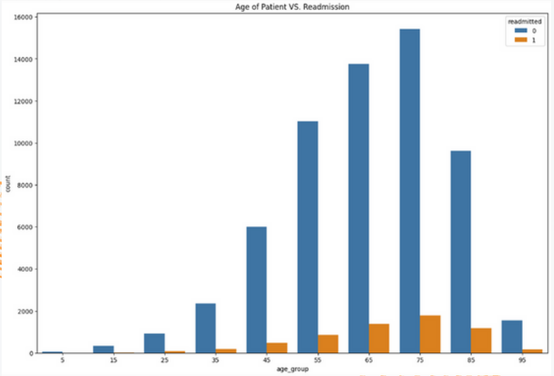
EXPLORATORY DATA ANALYSIS (EDA)

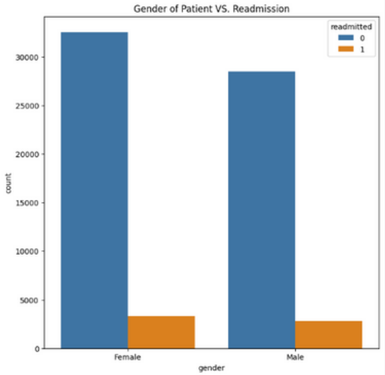


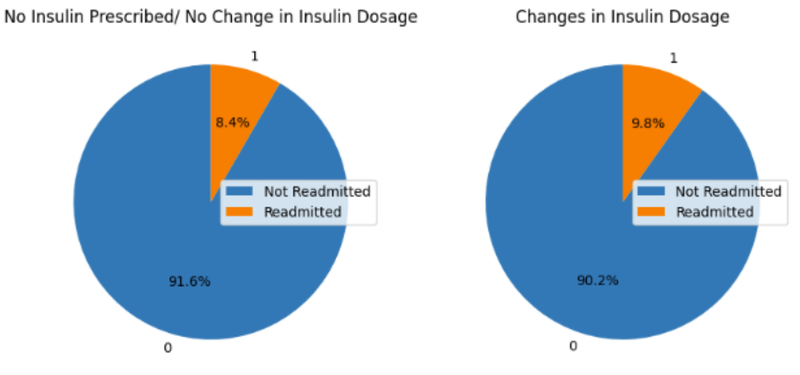




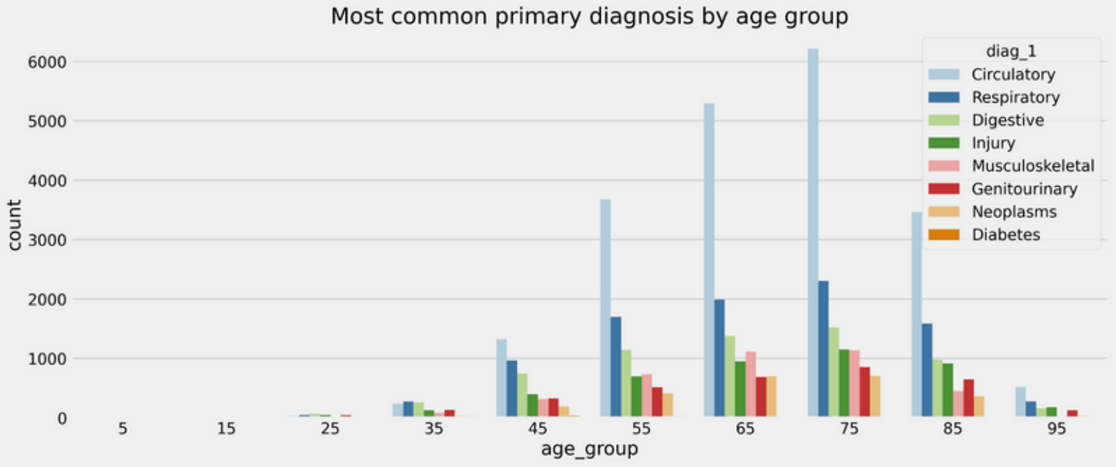


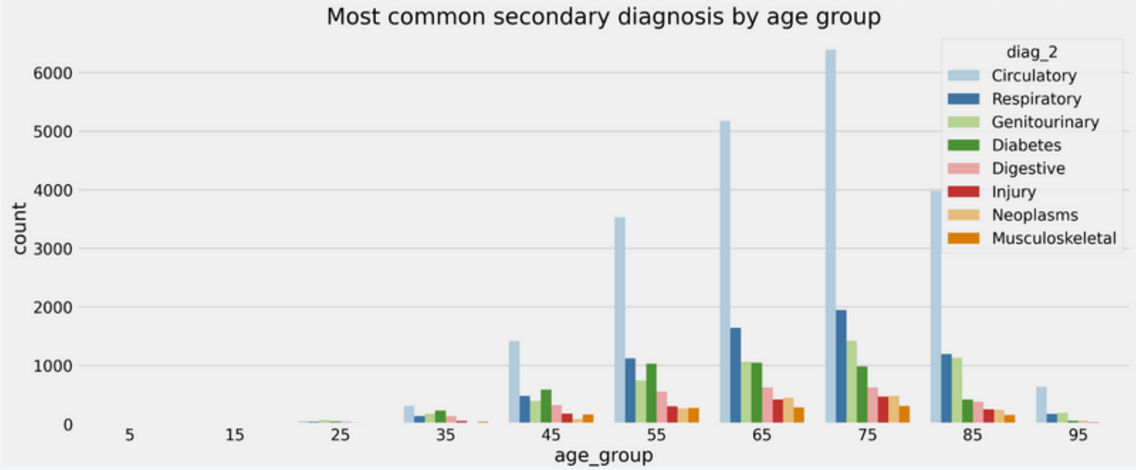


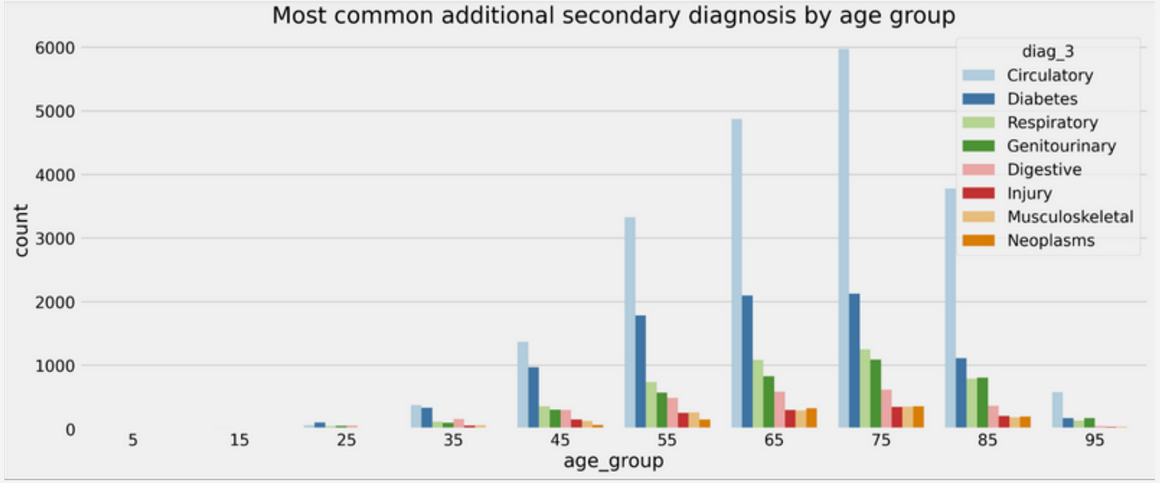




EDA


Most common primary diagnosis by age group
count
age_group
diag_1
Circulatory
Respiratory
Digestive
Injury
Musculoskeletal
Genitourinary
Neoplasms
Diabetes

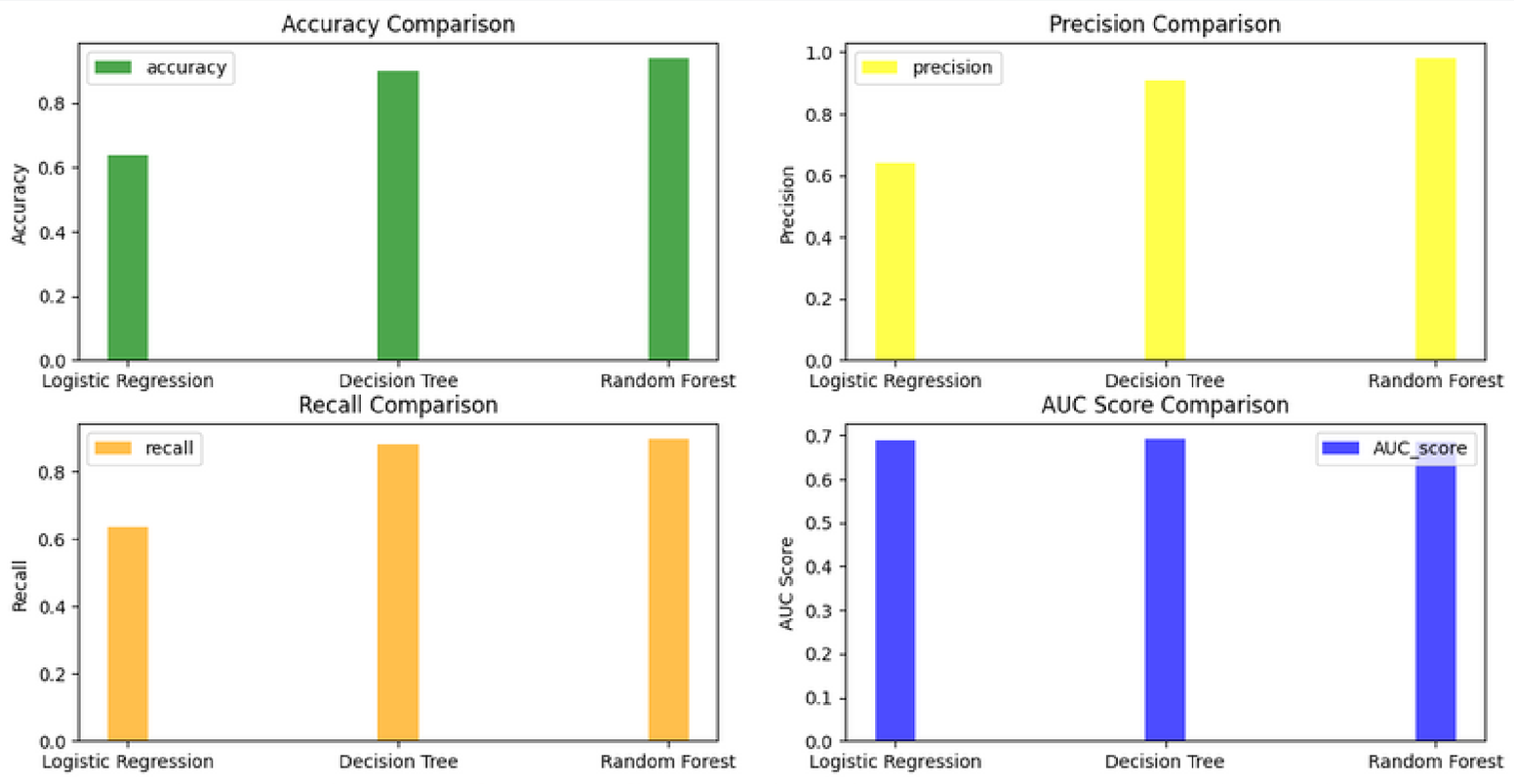

Most common secondary diagnosis by age group
count
age_group
diag_2
Circulatory
Respiratory
Genitourinary
Diabetes
Digestive
Injury
Neoplasms
Musculoskeletal


Most common additional secondary diagnosis by age group
count
age_group
diag_3
Circulatory
Diabetes
Respiratory
Genitourinary
Digestive
Injury
Musculoskeletal
Neoplasms

DELIVER

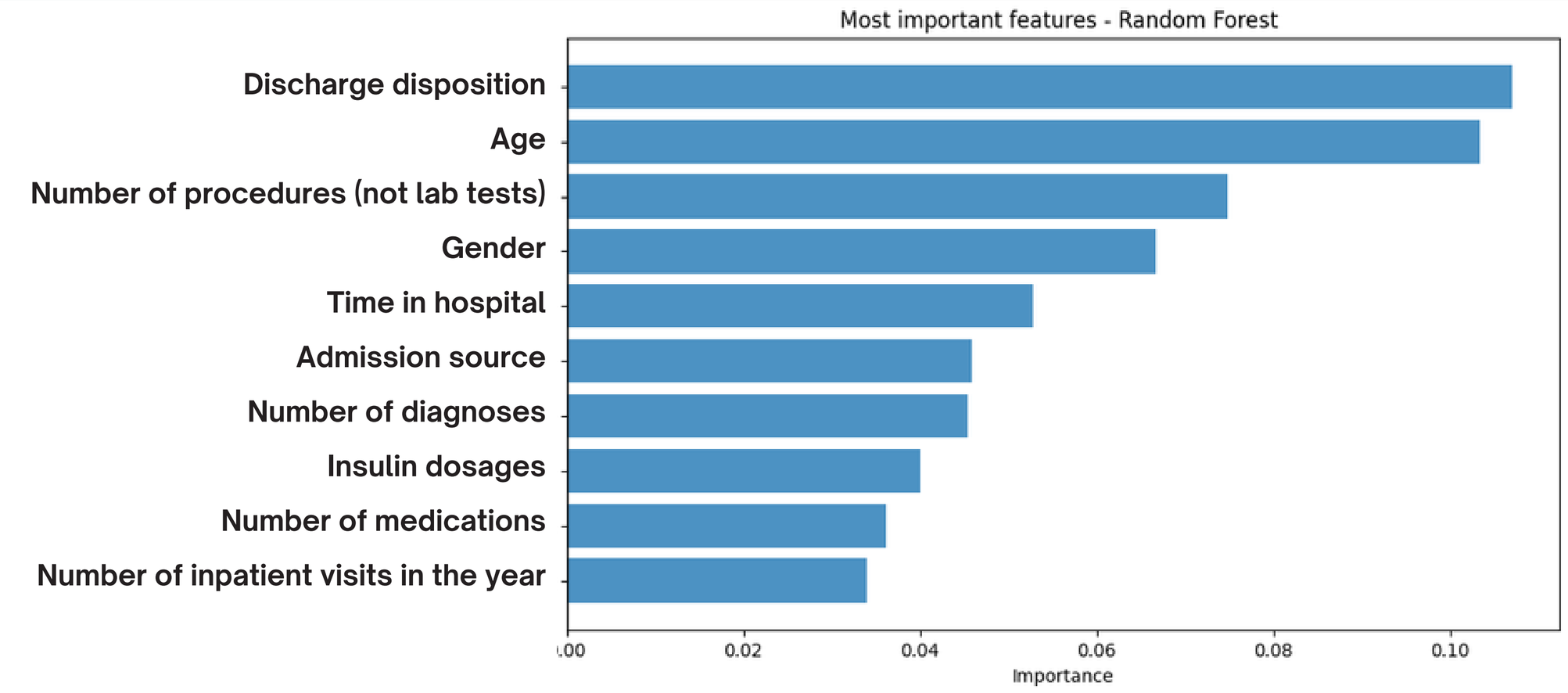
Institute of Data

MACHINE MODELS EVALUATION



- Overall, random forest has the highest accuracy, precision and recall scores compared to Logistic Regression and Decision Tree

TOP IMPORTANT FEATURES



CONCLUSIONS

This study utilized a comprehensive dataset that included demographic, clinical procedure-related, diagnostic-related, etc. for diabetic patients of all ages. The goal was to build a predictive model that could identify diabetic patients who are at a higher likelihood of being readmitted within 30 days.

Findings from this study offers numerous benefits in healthcare. Early identification of high-risk patients allows for timely interventions and improved patient outcomes. Optimal hospital resource allocation can be achieved. Population health management strategies can be proactively implemented based on predictive analytics. Data-driven decision making can inform providers, administrators, and policymakers in optimizing care processes. Quality improvement initiatives can be targeted based on readmission data analysis.

After analyzing three machine learning algorithms, namely Logistic Regression, Decision Tree, and Random Forest, it was determined that Random Forest performed the best with 94% accuracy rate. The study recommends that hospitals should focus not only on inpatient treatment but also on continued care after discharge.

FUTURE STUDY

Our results are promising and the proposed algorithms could be applied to other samples from different sourse hospitals and different countries. A model deployment can be built for health professionals to use and predict the probability of readmission for a certain patient by inputting patient's information such as demographic, discharge disposition, number of procedures taken during inpatient visit, etc.

Future work should also focus on applying other machine learning algorithms such as Support Vector Machines, Gradient Boosting Models, Deep Learning, Time-series Models.

Lastly, differences between shorter (less than 30 days) and longer (more than 30 days) readmission timeframe should be investigated as a criteria in future experiments.

BUSINESS VALUE

Assumptions

- Total number of diabetic patients in the U.S.: 30 millions (American Diabetes Association, 2022)
- The hospital readmission rate for U.S. diabetic patients: 18% (Ostling et al., 2017)
- The average cost of hospital readmission for a diabetic patient: US $10,000 (American Diabetes Association, 2018)
- This study aims to reduce the hospital readmission rate for U.S. diabetic patients by 9%

Cost of hospital readmissions without the predictive model

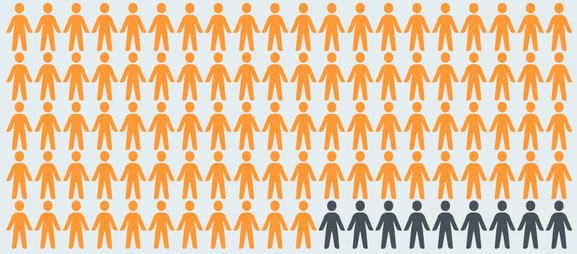
×

$10,000

=

54
billions

Cost of hospital readmissions with the predictive model



×

$10,000

=

27
billions

Cost savings

27
billions