DIABETES HOSPITAL READMISSION MODEL DOCUMENT

A Framework for Predictive Model on Hospital Readmission for Diabetic Patients

Abstract

Diabetes hospitalisation is associated with 13% of readmission rate, leading to increased cost on the Australian's health expenditure. This project aims to address this issue with the focus of finding associated factors to hospital readmission through data exploration and analysis. Machine learning is applied with the aim to create a model efficient for predicting readmission based on the chosen features. We use a dataset from UCI repository that contains 50 features and over 100,000 entries. Techniques including feature engineering, feature selection and hyperparameter tuning are applied in identifying the most relevant features and refinement of the final model. K-nearest neighbour was chosen as the final model. Through data analysis and modelling, it is suggested that the current data is insufficient in identifying readmission efficiently.

Li-Tin (Lily) Chen

j876023@gmail.com

Table of Contents

1 Purpose

This document is intended as a guideline for stakeholders and associated development team as reference to the data analysis and modelling process on the prediction of diabetes-related hospital readmission.

2 Project Description

Diabetes is one of the biggest challenges the Australia's health system is facing today. The total annual health expenditure on diabetes in Australia is estimated at a minimum of $6 billion. There were over 1 million diabetes-associated hospitalisations recorded in 2016-17, accounting for 10% of all hospitalisations in Australia. Moreover, 13% of the total diabetes hospitalisations were readmission within 30 days, resulting in huge burden and medical costs on Australian healthcare. Through data analysis, this project aims to identify main contributing factors of diabetes hospital readmission. A predictive model is created with the aim to predict readmission efficiently, based on the selected features. We aim to provide an insight in assisting inform discussion and policy decisions on the care management of diabetes in Australia.

3 Stakeholders

- Executive Director/Managing Director, Australian Institute of Health and Welfare

4 Business Question

Question: How to lower hospital readmission for diabetic patients to minimise unnecessary spending?
Current State:

- Average hospital admission cost per person: $7656
- Readmission rate: 13%

Business Value: Australia health fund can save 230 million on hospitalisation-related costs annually if able to prevent 3% readmissions (assumed that 3% of these readmissions are preventable).

5 Data Question

- What factors show strong contributions to the prediction of hospital readmission for diabetic patients?
- What model is the best at predict hospital readmission with the dataset? How is the model performance?

6 Dataset Description

Source: UCI repository

Data Set Description:

- Data Collection Timeframe 1999-2008
- 101,766 observations
- 50 Features including patient demographics, health conditions, laboratory tests, medications and care management during the recorded hospital encounter.

7 Data Science Process

7.1 Exploratory Data Analysis

Data Profiling – examining the characteristics and identifying issues with the raw dataset

- 50 columns
- 101766 rows
- Missing values in dataset
- Columns with incorrect data type

Data Cleaning – Making the data usable (preparing it for analysis)

- Drop irrelevant columns
- Drop columns with large numbers of missing values
- Drop columns with wide range of unique values
- Data type conversion
- Creating new column from combination of existing columns
- Rename columns

Final Data

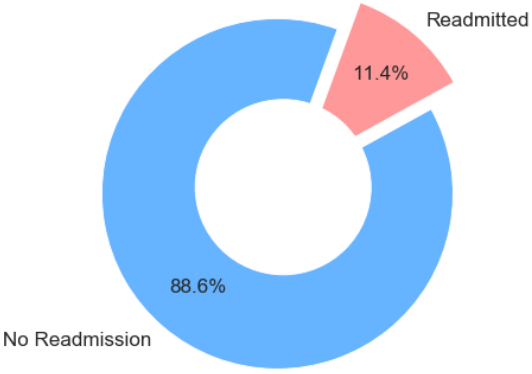
- Cleaned dataset: 97874 rows x 15 columns
- No missing values
- All data of numeric type
- Outcome variable: 'readmitted'

7.2 Data Directory:

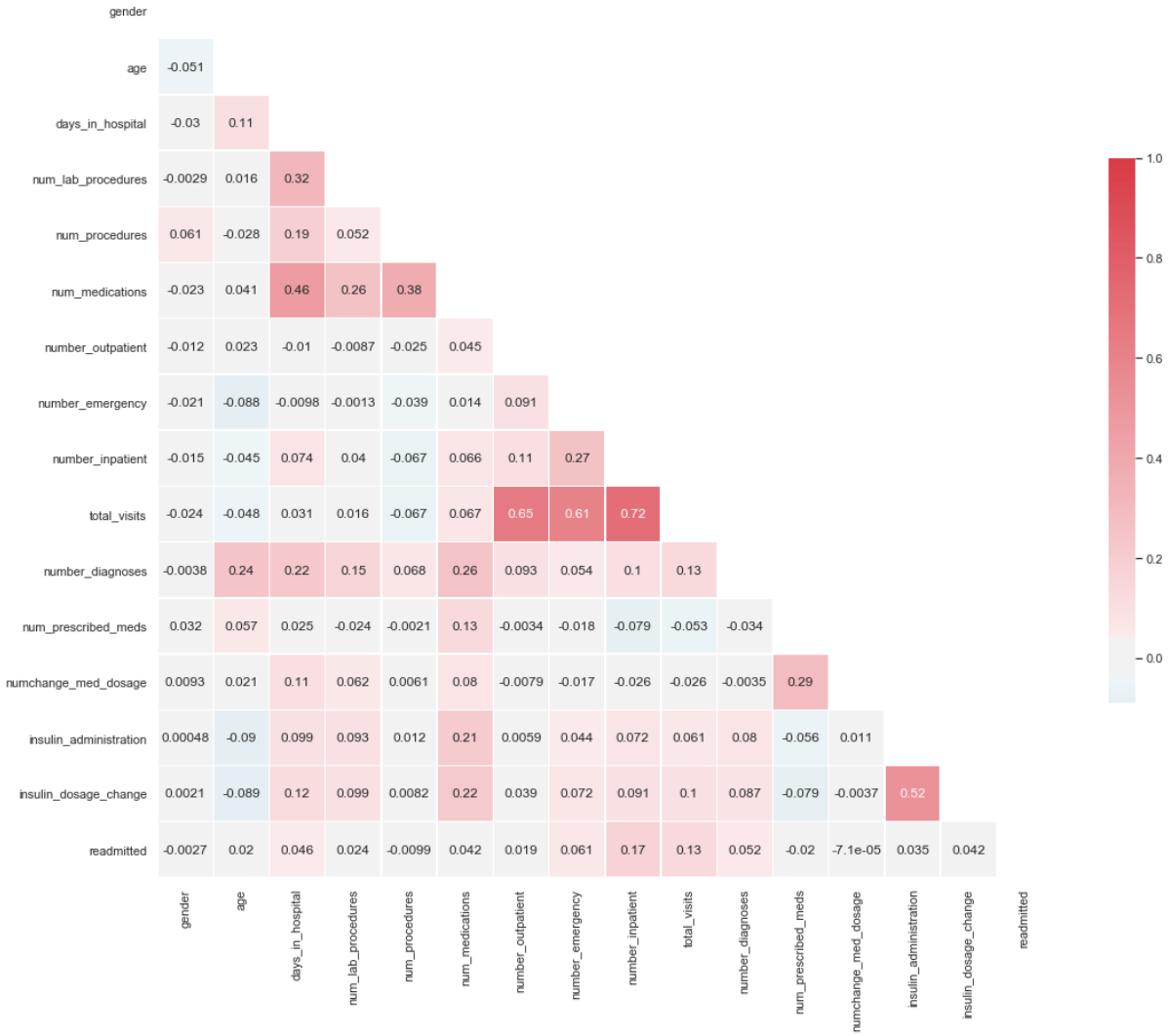
Column	Description
gender	0 = Female 1 = Male
age	Grouped in 10-year intervals: [0-10), [10-20), ..., [90-100)
days_in_hospital	Number of days between admission and discharge
num_lab_procedures	Number of lab tests performed during the encounter
num_procedures	Number of procedures (other than lab tests) performed during the encounter
num_medications	Number of distinct generic names administered during the encounter
number_outpatient	Number of outpatient visits of the patient in the year preceding the encounter
number_emergency	Number of emergency visits of the patient in the year preceding the encounter
number_inpatient	Number of inpatient visits of the patient in the year preceding the encounter
total_visits	Number of all hospital visits of the patient in the year preceding the encounter (number_outpatient + number_emergency + number_inpatient)
number_diagnoses	Number of diagnoses entered to the system
num_prescribed_meds	Number of diabetic medications prescribed
numchange_med_dosage	Number of dosage change of the prescribed diabetic medication
insulin_administration	0 = No Administration of Insulin 1 = Administration of Insulin
insulin_dosage_change	0 = No change of insulin dosage 1 = Change of insulin dosage
readmitted	0 = No readmission within 30days 1 = the patient was readmitted within 30 days after discharged

EDA Findings:

Outcome Variable Distribution: Imbalanced Dataset



Correlation Heatmap Between Features:



- Most features show little correlation with the target variable (readmitted)
- Top 5 correlated features: number_inpatient, total_visits, number_emergency, number_diagnoses and days_in_hospital

7.3 Data Balancing

The complete dataset is split into three subsets:

- Train set – for machine training purpose
- Test set – for model configuration and hyperparameters tuning
- Val set – for unbiased evaluation of the final model

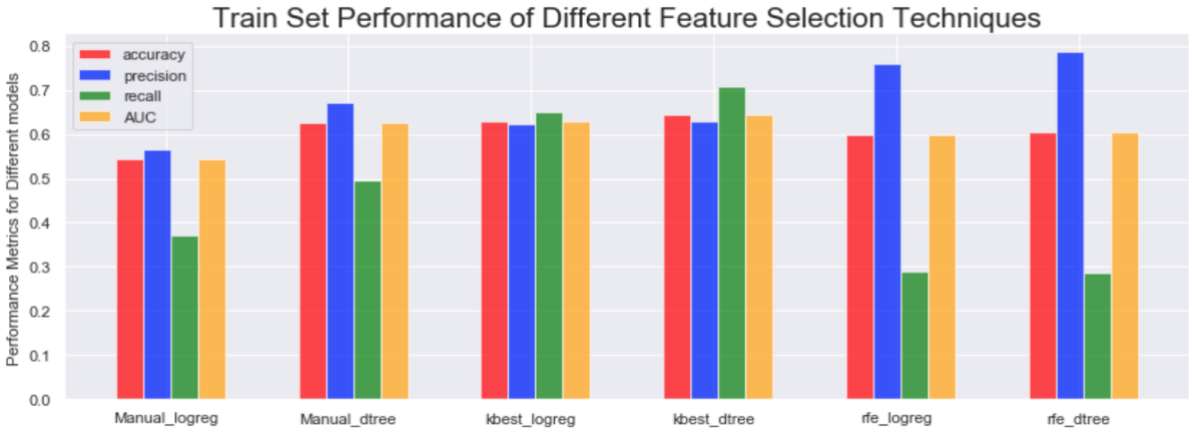
Data Balancing – imbalanced data can result in high biased model to majority class

- SMOTETomek technique is applied to the train set for data balancing
- SMOTETomek combines SMOTE (over-sampling) and Tomek Links (under-sampling) methods that address high bias issue commonly seen in oversampled data.
- Data balancing technique is only applied to the train set for model training purpose; test and validation sets did not get balance as it can affect the real-world problem thus resulting in over-optimistic outcome.
- Reference: https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.combine.SMOTETomek.html

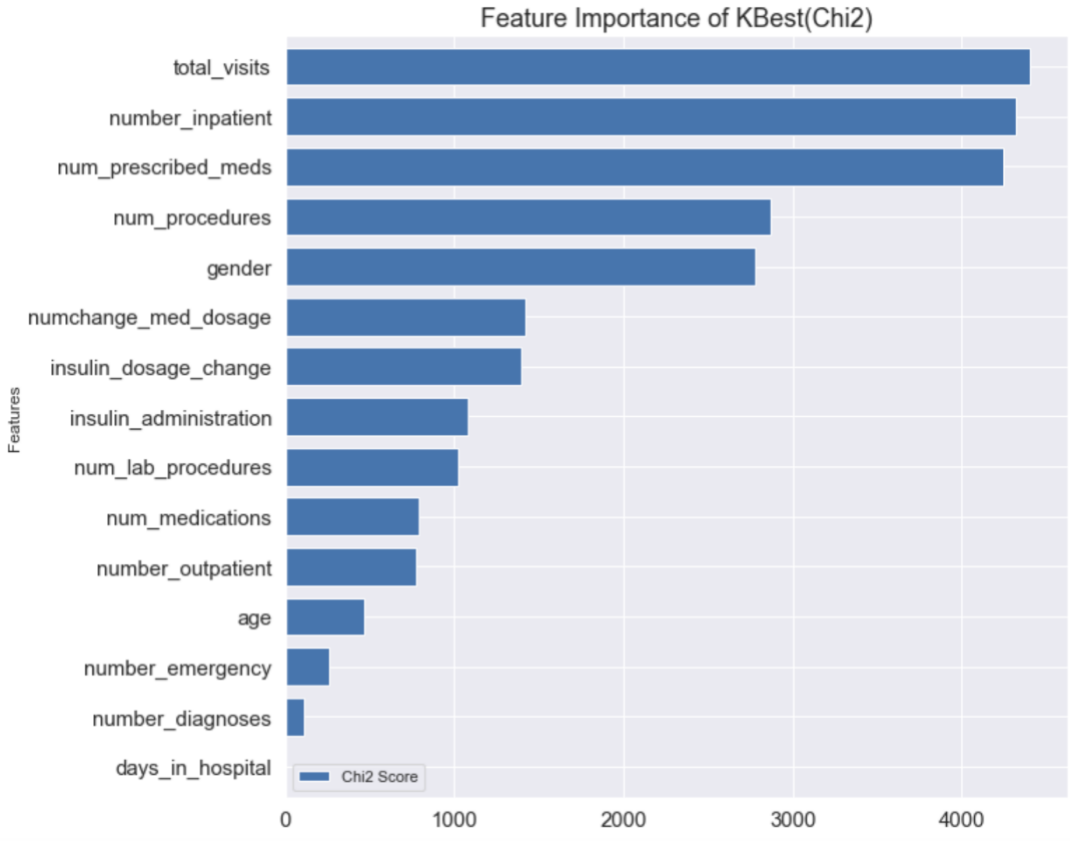
7.4 Feature Selection

Feature Selection – identify strong predictors of the outcome

- Three feature selection techniques are applied on two models (logistic regression and decision tree)
    - Based on correlation heatmap: relationship between feature and outcome variables
    - SelectKBest based on chi2 score: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html
    - Recursive feature elimination: recursively removes the weakest features until the specified number of features is reached https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html
- SelectKBest is chosen as the final feature selection technique
    - highest recall
    - least variability across the four performance metrics
- The top 5 features with the highest chi2 score (Chi2 > 2000) are selected as final features for modelling.
    - total_visits, number_inpatient, num_prescribed_meds, num_procedures, gender


Train Set Performance of Different Feature Selection Techniques
Performance Metrics for Different models
accuracy
precision
recall
AUC
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Manual_logreg
Manual_dtree
kbest_logreg
kbest_dtree
rfe_logreg
rfe_dtree


Test Set Performance of Different Feature Selection Techniques
Performance Metrics for Different models
accuracy
precision
recall
AUC
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Manual_logreg
Manual_dtree
kbest_logreg
kbest_dtree
rfe_logreg
rfe_dtree


Feature Importance of KBest(Chi2)
Features
total_visits
number_inpatient
num_prescribed_meds
num_procedures
gender
numchange_med_dosage
insulin_dosage_change
insulin_administration
num_lab_procedures
num_medications
number_outpatient
age
number_emergency
number_diagnoses
days_in_hospital
Chi2 Score
0
1000
2000
3000
4000

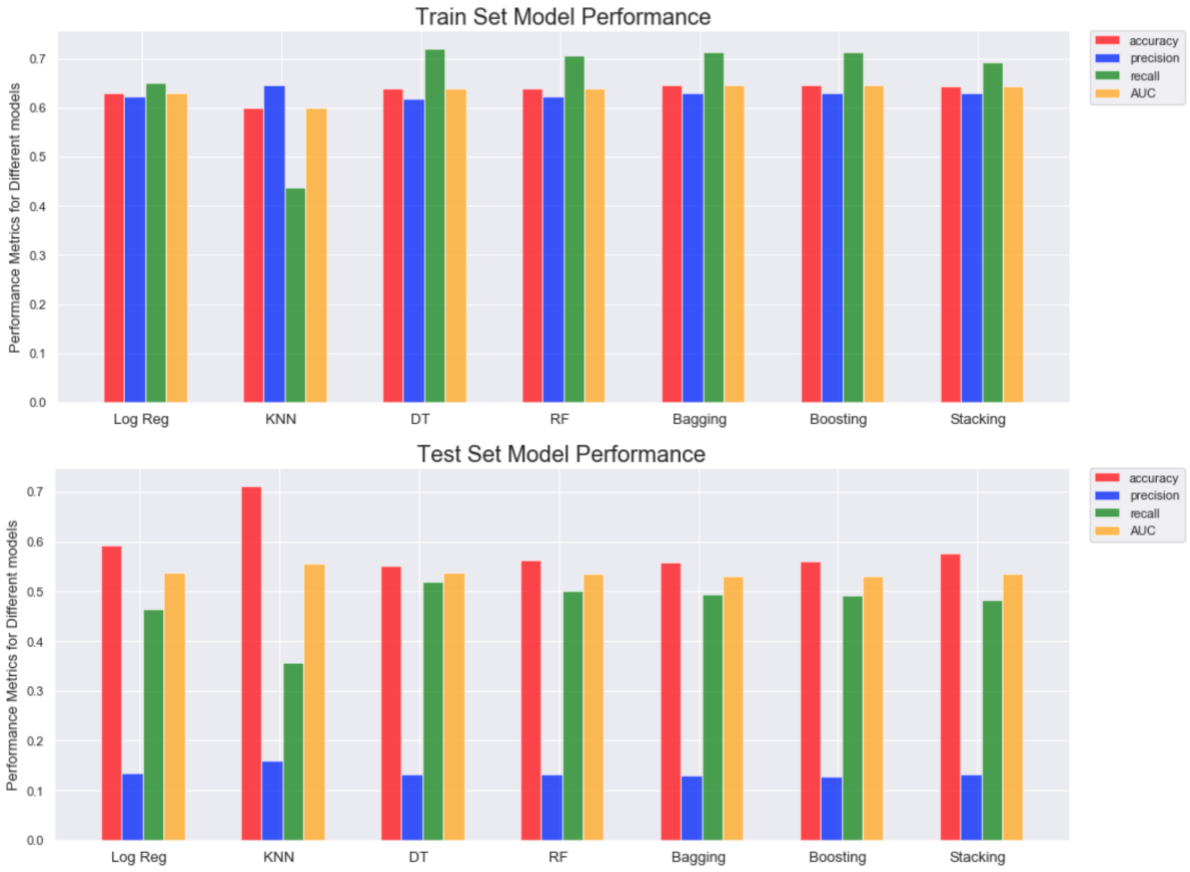
7.5 Models Comparison

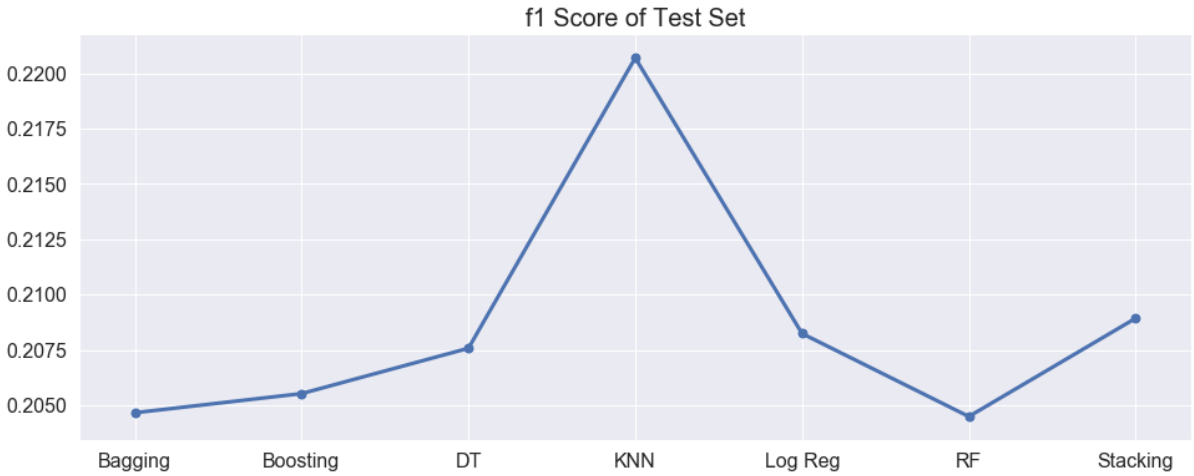
Hyperparameter Tuning

- Random search is a technique used for finding random combinations of hyperparameters that results in the best solution for the selected model.
- RandomizedSearchCV is applied on three base models: Logistic Regression, K-Nearest Neighbors and Decision Tree
- Search time: <4mins

Training Models

- 7 Models: Logistic Regression, K-Nearest Neighbors, Decision Tree Classifier, Random Forest, Bagging, Boosting and Stacking
- Hyper-parameters where chosen using randomized search with 5-fold cross validation
- Training time under 5mins





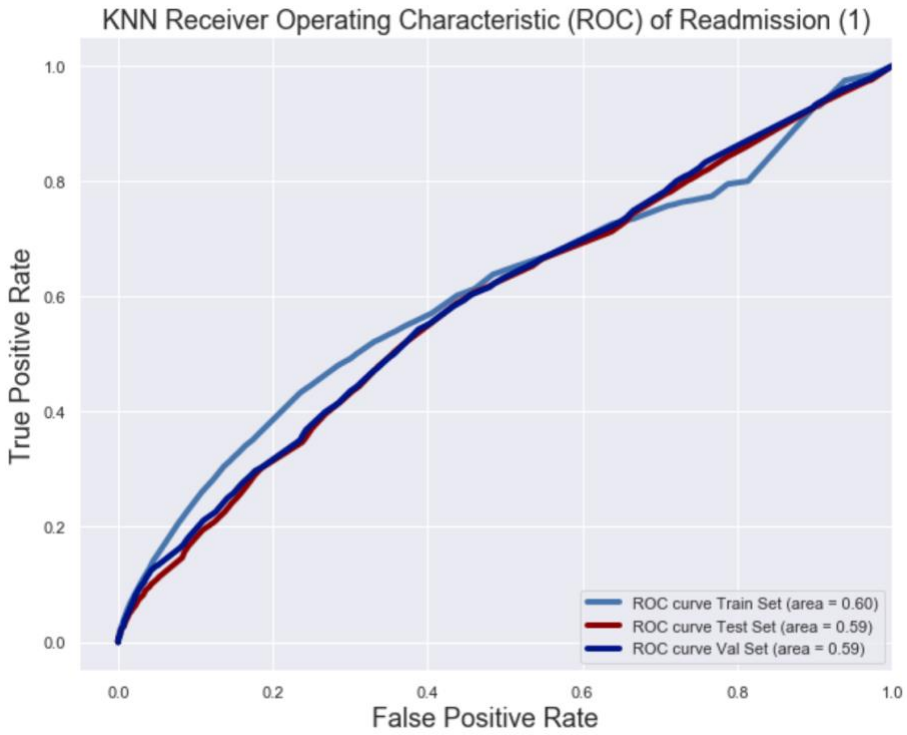
	Model	Recall	Precision	f1 Score
0	Log Reg	0.465	0.134	0.208
1	KNN	0.356	0.160	0.221
2	DT	0.497	0.130	0.206
3	RF	0.504	0.132	0.209
4	Bagging	0.495	0.130	0.205
5	Boosting	0.498	0.130	0.206
6	Stacking	0.480	0.132	0.208

Findings

- DT, RF, Bagging, Boosting and Stacking show similar result with very minor difference in the recall rate.
- Based on the graphs above, KNN method is chosen as the best model.
  - least overfitting compared to other models
  - highest f1 score 0.22: a measure of model accuracy that considers both the precision and the recall of the model
    - precision 0.16: percentage of predicted values that are relevant
    - recall 0.36: percentage of total relevant results that are correctly classified by the model

7.6 Final Model Evaluation

- Final model is applied to val set for unbiased evaluation



	Train Set	Test Set	Val Set
AUC	0.599	0.556	0.560
Recall	0.437	0.356	0.359
Precision	0.646	0.160	0.166
f1 Score	0.521	0.221	0.227

Final Model Evaluation

- Similar ROC curve between train, test and val sets
- Similar performances between test and val sets, suggesting that the model fits 'unseen' data as expected.
- Based on recall, final model is able to predict 35% of actual readmission.
- Based on precision, 16% of the predicted readmissions are actual readmissions.
- f1 score ranges between 0.22 and 0.23

7.7 Outcome Summary

- The top 5 strongest predictors were selected with Chi2 score using SelectKBest
- KNN is selected as the best model out of the seven trained models
    - AUC of 0.56 – tells how well the model is able to classify readmission, which is almost like a random classifier (AUC of 0.5)
    - Recall value of 0.36 – in reality, only 36% of actual readmissions were labelled correctly
    - Precision value of 0.16 – only 16% out of all predicted readmissions were actually readmitted
    - F1 score of 0.22 – harmonic mean between precision and recall.
- Several techniques were used in finding most relevant factors and model refinement. However, the model performance still does not meet the satisfactory standard for model deployment.
- Model is incapable in accurately predicting most of the readmissions due to low correlation between the features and the outcome variable.
- Factors that we thought might be useful in the prediction of readmission (prior to data analysis) are found to give low predictive values.

8 Recommendations

Project findings show that the current dataset on diabetes readmission provide inadequate statistical information on diabetes hospital readmission. Further investigation and data collection are required in order to provide sufficient information for model development and refinement.

Statistically, type 2 diabetes account for more than 85% of the total number of diagnosed diabetes in Australia. it is associated with several hereditary factors and lifestyle risk factors such as poor diet, inadequate physical activity, being overweight or obese. These risk factors were not included in the dataset but may provide values in the identification of predictive factors of readmission. Other possible factors such as poor medication adherence has also been found in past research to be associated with increased hospitalisation and complications. This may be related with inadequate discharge educations and poor patient understanding. Investigation in these factors may also be relevant to the prediction of hospital readmission for diabetic patients.

9 References

Codes:

- Exploratory Data Analysis: Lily-Chen/capstone/code/Lily Capstone Project Diabetes Hospital Readmission EDA.ipynb
- Modelling: Lily-Chen/capstone/code/Lily Capstone Project Diabetes Hospital Readmission Modelling FINAL.ipynb

Websites:

- Data Source: https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008#
- https://static.diabetesaustralia.com.au/s/fileassets/diabetes-australia/e7282521-472b-4313-b18e-be84c3d5d907.pdf
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4966497/#:~:text=Poor%20medication%20adherence%20in%20T2D,and%20managing%20complications%20of%20diabetes.
- https://towardsdatascience.com/hyper-parameter-tuning-and-model-selection-like-a-movie-star-a884b8ee8d68
- https://www.aihw.gov.au/reports/diabetes/diabetes/contents/what-is-diabetes
- https://www.sciencedirect.com/science/article/pii/S1059131117302455
- https://link.springer.com/article/10.1186/s12913-018-3723-4
- https://lukesingham.com/whos-going-to-leave-next/
- https://towardsdatascience.com/the-5-classification-evaluation-metrics-you-must-know-aa97784ff226
- https://link.springer.com/article/10.1007/s11892-018-0989-1
- https://towardsdatascience.com/hackcvilleds-4636c6c1ba53

Python Library:

- Numpy
- Pandas
- Matplotlib
- Seaborn
- Imblearn
- Sklearn
- Scipy
- Time
- IPython