CASE STUDY
READMISSION

HOANG NHA NGUYEN
(671491808)

WHY ?

- Hospital readmission is a high-priority health care quality measure and target for cost reduction, particularly within 30 days of discharge (30-day readmission is typically concerning).
- Reducing readmission rates among patients with diabetes has the potential to greatly reduce health care costs while simultaneously improving care.

Workflow

01	**Database Selection**	• Dataset - diabetic readmission_patients.csv • Target variable: ‘readmission’
02	**Data Pre-processing**	• Exploratory Data Analysis • Data cleaning • Feature Engineering - One hot coding
03	**Data Modelling**	• Logistic Regression
04	**End Result**	• Models Evaluation

Goal

Predicting 30-day readmission of patients with diabetes using data on prior medical encounters and patients historic clinical data.

Develop a machine learning model that can predict the readmission of patients with diabetes.

Dataset

The dataset includes multiple encounters of over 70 thousand diabetes patients. It contains 101766 rows and 50 features including:

- patient demographics
- patient characteristics
- medical history
- drugs
- previous admission related information.

Data Pre-processing

- No null values , but data labeled as “?” are considered missing data
- After transpose, **192849 missing values** are detected in the following columns:
  - **diag_1** – 21
  - **medical_specialty** – 49949 *
  - **diag_3** – 1423
  - **weight** – 98569 *
  - **diag_2** – 358
  - **payer_code** – 40256 *
  - **race** – 2273

- Remove rows with missing values for “**diag_1**”, “**diag_3**”, “**diag_2**”, and “**race**”
- Remove the following variables:
  - “**medical_specialty**”, “**weight**”, and “**payer_code**” -> high % of missing values
  - “**examide**” and “**citoglipton**” because they have only 1 unique value -> not useful

Data-Preprocessing (cont)

- The target variable - "**readmission**" is reduced into 2 classes
  - 1- readmission within 30 days
  - 0 - no readmission within 30 days to indicate whether the patient had a readmission within 30 days of the previous encounter
- **Service_utilization** - sum of "outpatient"," emergency", and "inpatient" admission values
- Values for the feature "**age**" are grouped into three categories –
  - "<50"
  - "50-70"
  - "70+"
- Convert all drug related columns into binary classes (Y/N)
- **"Drug_change"** - new feature that captures the total number of medications that were changed (up/down) in an encounter. It is the total count of all the drugs whose dosage were modified (either up or down) in the encounter

Data Pre-processing (cont)

- "**A1C result**" values are translated to indicate levels
- "**Admission type**", "**Admission source**", and "**Discharge Disposition Ids**" are mapped into their respective categories as mentioned in the data dictionary - using SWITCH tool to map:

admission_type_id (9)	[1,2] -> 'Emergency', [3]' -> 'Elective', [4,7] -> 'other', [5,6,8] -> 'Unknown'
discharge_disposition_id (29)	[6] -> 'home', [2,3,4,5,15,16,17,,22,23,24,27,28] ->'Another Facility', [8,13,14] -> 'Nursing Home', [7,9,10,11,12,19,20] -> ''Other', [18,25] -> 'Unknown'
admission_source_id (21)	[1,2,3] -> 'Referral, [4,5,6,10,22,25] -> 'Transfer', [7] -> 'ER', [8,11,13,14] -> 'other', [9,17,20] -> 'Unknown'
change (2)	1 -> "Yes" , 0 -> "No" - if there was a change in diabether medications
gender (2)	1 -> Male, 0 -> Female
A1Cresult	"none" if not measured, ">7%" - > "Pre-diab", ">8%" -> "Diab", "Norm" -> "Normal"

Data Pre-processing (cont)

- The primary diagnosis codes are used for modeling purposes i.e. **diag_1** (923 distinct values)
- **“level1_diag1”** - if contains “V” and “E” -> 0, else -> “diag_1”

Diagnosis Codes	Group Name	ID	Description
390–459, 785	Circulatory	1	Diseases of the circulatory system
460–519, 786	Respiratory	2	Diseases of the respiratory system
520–579, 787	Digestive	3	Diseases of the digestive system
250.xx	Diabetes	4	Diabetes mellitus
800–999	Injury	5	Injury and poisoning
710–739	Musculoskeletal	6	Diseases of the musculoskeletal system and connective tissue
580–629, 788	Genitourinary	7	Diseases of the genitourinary system
140–239	Neoplasms	8	Neoplasms
780, 781, 784, 790–799	Other	0	Other symptoms, signs, and ill-defined conditions
240–279, without 250			Endocrine, nutritional, and metabolic diseases and immunity disorders, without diabetes
680–709, 782			Diseases of the skin and subcutaneous tissue
001–139			Infectious and parasitic diseases
290–319			Mental disorders
E–V			External causes of injury and supplemental classification
280–289			Diseases of the blood and blood-forming organs
320–359			Diseases of the nervous system
630–679			Complications of pregnancy, childbirth, and the puerperium
360–389			Diseases of the sense organs
740–759			Congenital anomalies

Final Dataset Review - Variable Selection

- Multiple encounters of patients are removed. Only the first encounter of each patient is retained -> **Unique Identifiers** only
- Remove “encounter_id”, “patient_nbr”
- Remove “number_outpatients”, “number_inpatients”, “number_emergency” -> **“service_utilization”**
- Remove “diag_1”, “diag_2”, and “diag_3” -> **“level1_diag1”**
- **Final dataset has 67580 encounters of patients with 40 features**

Feature Engineering - One hot encoding

- Apply 1`hot encoding to all categorical variables
- Drop 1 category from each categorical variable.
  67580 x 63 columns (after hot encoding)
- Change the data types of the following variables from Boolean (T/F) to Int16 (0/1)

Model 1 - Logistic Regression

- Only 10% of patients get re-admitted

  -> Apply oversampling technique

Coefficients: (7 not defined because of singularities)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.325285	3.145e-01	-4.21445	3e-05	***
gender	0.031886	4.660e-02	0.68426	0.49381	
time_in_hospital	0.036795	9.189e-03	4.00408	6e-05	***
num_lab_procedures	0.004209	1.407e-03	2.99056	0.00278	**
num_procedures	-0.047652	1.576e-02	-3.02384	0.0025	**
num_medications	0.006083	3.745e-03	1.62417	0.10434	
number_diagnoses	0.060561	1.382e-02	4.38115	1e-05	***
metformin	-0.214324	6.656e-02	-3.22023	0.00128	**
repaglinide	0.123033	1.808e-01	0.68034	0.49629	
nateglinide	0.135069	2.936e-01	0.46001	0.64551	
chlorpropamide	-1.893681	1.067e+00	-1.77539	0.07583	.
glimepiride	-0.057935	1.093e-01	-0.52987	0.59621	
glipizide	0.018402	7.571e-02	0.24307	0.80795	
glyburide	-0.094936	8.305e-02	-1.14318	0.25296	
tolbutamide	11.453122	1.970e+02	0.05815	0.95363	
pioglitazone	-0.193237	9.520e-02	-2.02988	0.04237	*
rosiglitazone	-0.156632	9.784e-02	-1.60087	0.1094	
acarbose	-0.318255	4.574e-01	-0.69582	0.48654	
miglitol	0.489315	1.431e+00	0.34187	0.73245	
tolazamide	-11.348362	1.970e+02	-0.05762	0.95406	
insulin	-0.012356	8.238e-02	-0.15000	0.88076	
glyburide.metformin	-0.022711	3.089e-01	-0.07351	0.9414	
change	0.059535	7.731e-02	0.77010	0.44124	

Model 1 - Logistic Regression

diabetesMed	0.264786	6.717e-02	3.94195	8e-05	***
service_utilization	0.208730	1.539e-02	13.56248	< 2.2e-16	***
race_AfricanAmerican	0.253325	2.070e-01	1.22392	0.22098	
race_Asian	0.069084	3.217e-01	0.21476	0.82995	
race_Caucasian	0.299101	2.013e-01	1.48574	0.13735	
race_Hispanic	0.256638	2.539e-01	1.01082	0.3121	
age_50.70	0.088426	7.648e-02	1.15625	0.24758	
age_70.	0.309214	7.720e-02	4.00532	6e-05	***
admission_type_id_	-0.715774	1.599e-01	-4.47537	1e-05	***
admission_type_id_Elective	-0.198197	1.282e-01	-1.54595	0.12212	
admission_type_id_Emergency	-0.155980	1.124e-01	-1.38735	0.16533	
admission_type_id_Other	-0.534120	1.249e+00	-0.42778	0.66881	
admission_source_id_ER	0.024339	1.231e-01	0.19778	0.84322	
admission_source_id_Other	11.809874	1.970e+02	0.05996	0.95219	
admission_source_id_Referral	0.040911	1.259e-01	0.32486	0.74528	
admission_source_id_Transfer	0.005323	1.522e-01	0.03499	0.97209	
max_glu_serum_.200	-0.238621	2.450e-01	-0.97394	0.33009	
max_glu_serum_.300	-0.419133	2.673e-01	-1.56791	0.1169	
max_glu_serum_None	-0.549293	1.792e-01	-3.06446	0.00218	**
A1Cresult_Diab	-0.131641	1.415e-01	-0.93002	0.35236	
A1Cresult_None	0.030085	1.188e-01	0.25331	0.80003	
A1Cresult_Normal	-0.215684	1.535e-01	-1.40509	0.15999	
level1_diag1_Circulatrory	0.403022	7.839e-02	5.14095	2.73e-07	***
level1_diag1_Diabetes	0.507180	1.064e-01	4.76594	1.87e-06	***
level1_diag1_Digestive	0.201850	1.026e-01	1.96691	0.04919	*
level1_diag1_Genitourinary	0.239700	1.201e-01	1.99518	0.04602	*
level1_diag1_Injury	0.575755	1.079e-01	5.33404	9.60e-08	***
level1_diag1_Musculoskeletal	0.377772	1.239e-01	3.04888	0.0023	**
level1_diag1_Neoplasms	0.455539	1.406e-01	3.24109	0.00119	**

Model 1 - Logistic Regression

level1_diag1_Neoplasms	0.455539	1.406e-01	3.24109	0.00119	**
level1_diag1_Others	0.294874	8.585e-02	3.43485	0.00059	***
race_Other	NA	NA	NA	NA	
age_.50	NA	NA	NA	NA	
admission_type_id_Unknown	NA	NA	NA	NA	
admission_source_id_Unknown	NA	NA	NA	NA	
max_glu_serum_Norm	NA	NA	NA	NA	
A1Cresult_Pre.diab	NA	NA	NA	NA	
level1_diag1_Respiratory	NA	NA	NA	NA	

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

Null deviance: 11275 on 8139 degrees of freedom
Residual deviance: 10716 on 8080 degrees of freedom
McFadden R-Squared: 0.04958, Akaike Information Criterion 10822

Model 1 - Logistic Regression

	Actual Positive	Actual Negative
Predicted Positive	2610 (61%)	1672 (39%)
Predicted Negative	1474 (38.3%)	2377 (61.7%)

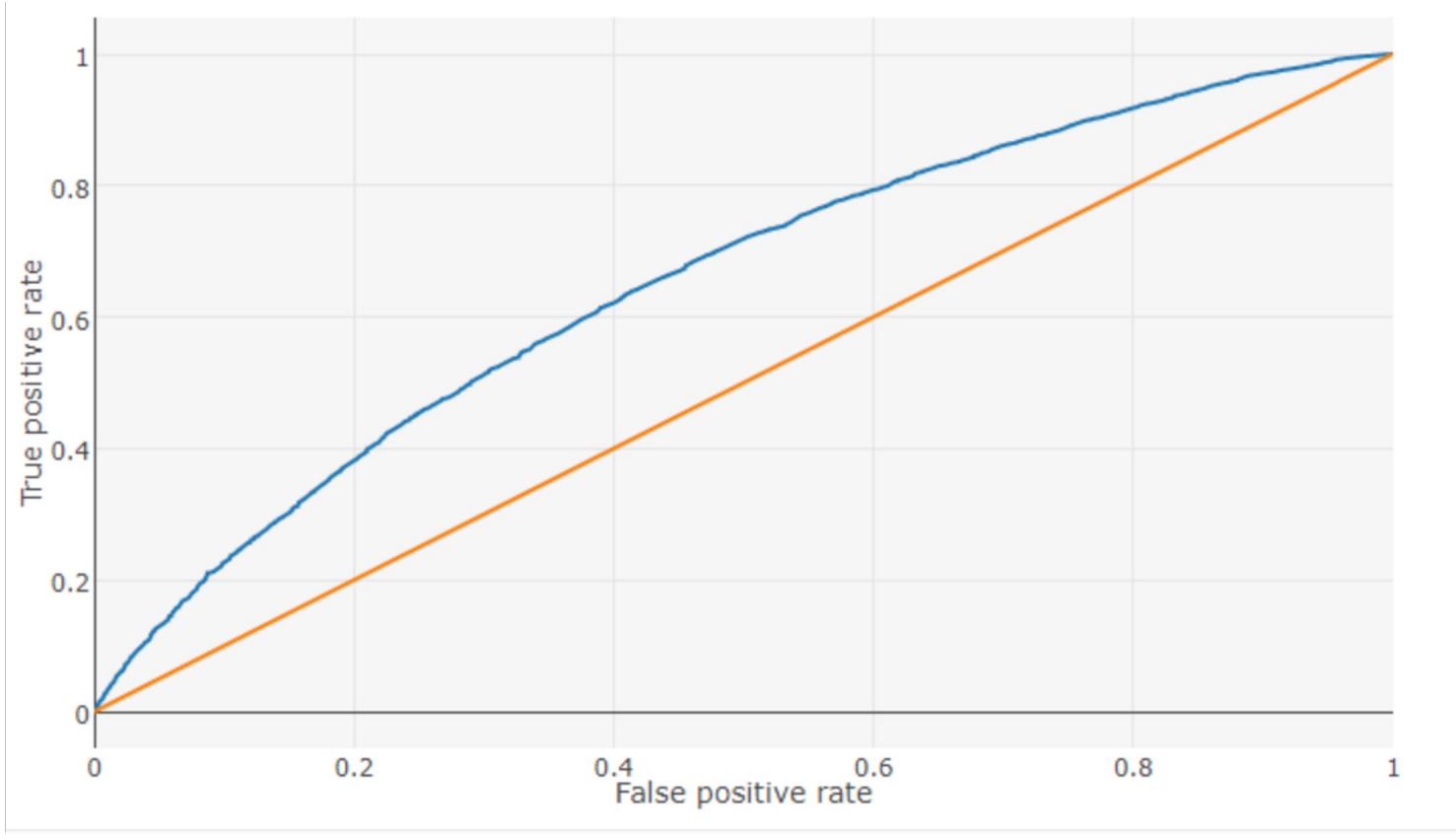
ACCURACY
0.613

PRECISION
0.61

RECALL
0.639

F1
0.624

OPTIMAL PROBABILITY CUTOFF
0.483



Model 2 - Decision Tree

Model Summary

Variables actually used in tree construction:

[1] age_70. diabetesMed metformin num_lab_procedures race_Caucasian service_utilization time_in_hospital

Root node error: 4049/8133 = 0.49785

n= 8133

Pruning Table

Level	CP	Num Splits	Rel Error	X Error	X Std Dev
1	0.1877007	0	1.00000	1.0333	0.011132
2	0.0072857	1	0.81230	0.8123	0.010931
3	0.0034576	4	0.78834	0.8081	0.010922
4	0.0027167	9	0.77031	0.8044	0.010914

Model 2 - Decision Tree

	Actual Positive	Actual Negative
Predicted Positive	2390 (62.6%)	1425 (37.4%)
Predicted Negative	1694 (39.2%)	2624 (60.8%)

Leaf Summary

node), split, n, loss, yval, (yprob)
* denotes terminal node

1) root 8133 4049 1 (0.4978483 0.5021517)
2) service_utilization< 0.5 4722 1981 0 (0.5804744 0.4195256)
4) time_in_hospital< 3.5 2213 807 0 (0.6353366 0.3646634) *
5) time_in_hospital>=3.5 2509 1174 0 (0.5320845 0.4679155)
10) age_70.< 0.5 1224 502 0 (0.5898693 0.4101307) *
11) age_70.>=0.5 1285 613 1 (0.4770428 0.5229572)
22) race_Caucasian< 0.5 232 97 0 (0.5818966 0.4181034) *
23) race_Caucasian>=0.5 1053 478 1 (0.4539411 0.5460589)
46) diabetesMed< 0.5 242 114 0 (0.5289256 0.4710744) *
47) diabetesMed>=0.5 811 350 1 (0.4315660 0.5684340) *
3) service_utilization>=0.5 3411 1308 1 (0.3834653 0.6165347)
6) service_utilization< 3.5 2747 1124 1 (0.4091736 0.5908264)
12) num_lab_procedures< 32.5 667 333 0 (0.5007496 0.4992504)
24) metformin>=0.5 148 53 0 (0.6418919 0.3581081) *
25) metformin< 0.5 519 239 1 (0.4605010 0.5394990)
50) age_70.< 0.5 259 121 0 (0.5328185 0.4671815) *
51) age_70.>=0.5 260 101 1 (0.3884615 0.6115385) *
13) num_lab_procedures>=32.5 2080 790 1 (0.3798077 0.6201923) *
7) service_utilization>=3.5 664 184 1 (0.2771084 0.7228916) *

Accuracy 61.7 %
Proportion of correct predictions in the data

F1 Score 60.5 %
Harmonic mean of Recall and Precision

Precision 62.6 %
Proportion of values predicted positive, that were actually positive

Recall 58.5 %
Proportion of values actually positive, that were predicted positive


service_utilization
< 0.5
time_in_hospital
< 3.5
service_utilization
< 3.5
0
age_70.
< 0.5
num_lab_procedures
< 32.5
1
0
race_Caucasian
< 0.5
metformin
>=0.5
1
0
diabetesMed
< 0.5
0
age_70.
< 0.5
0
1
0
1

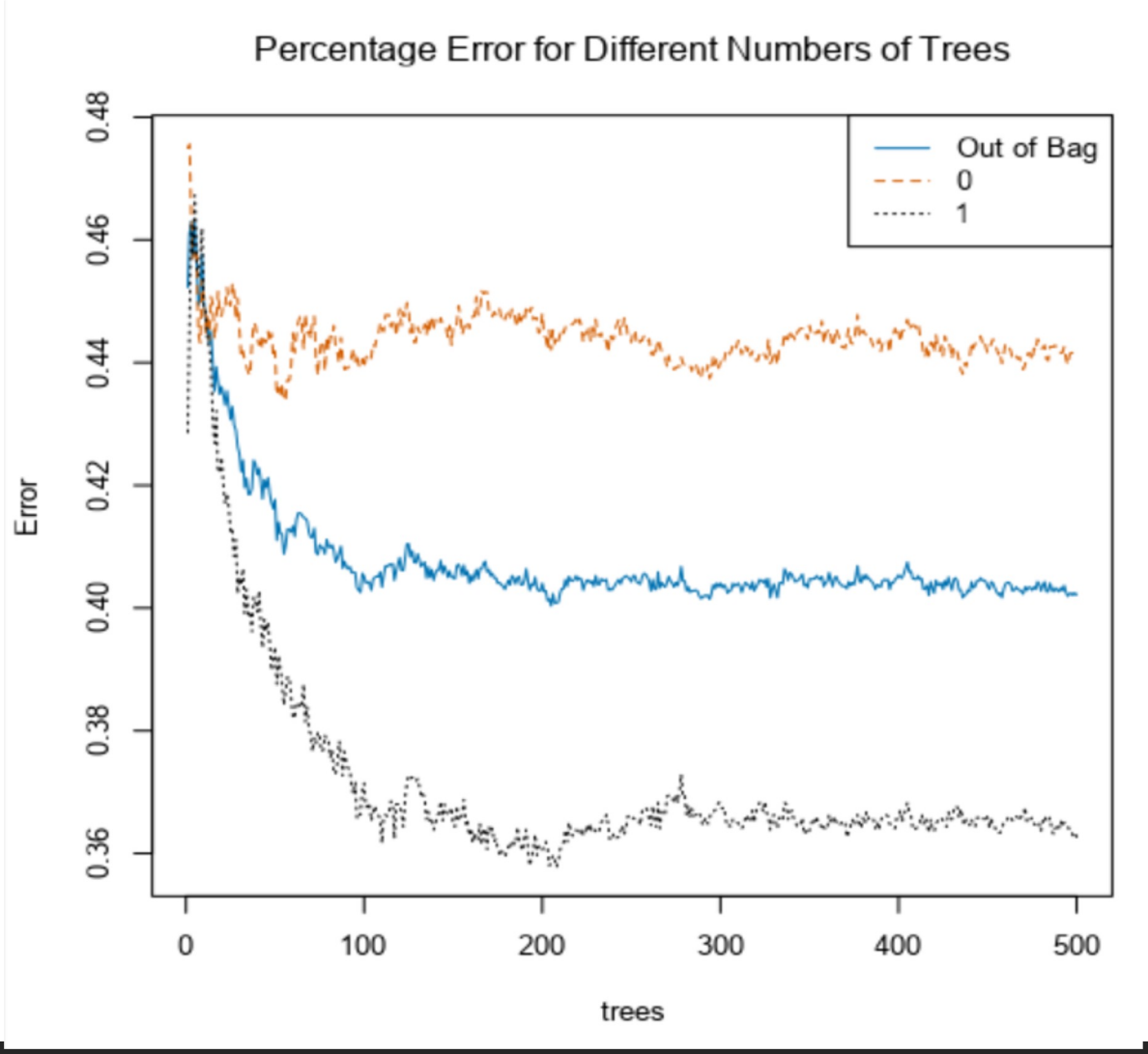
Model 3 - Random Forest

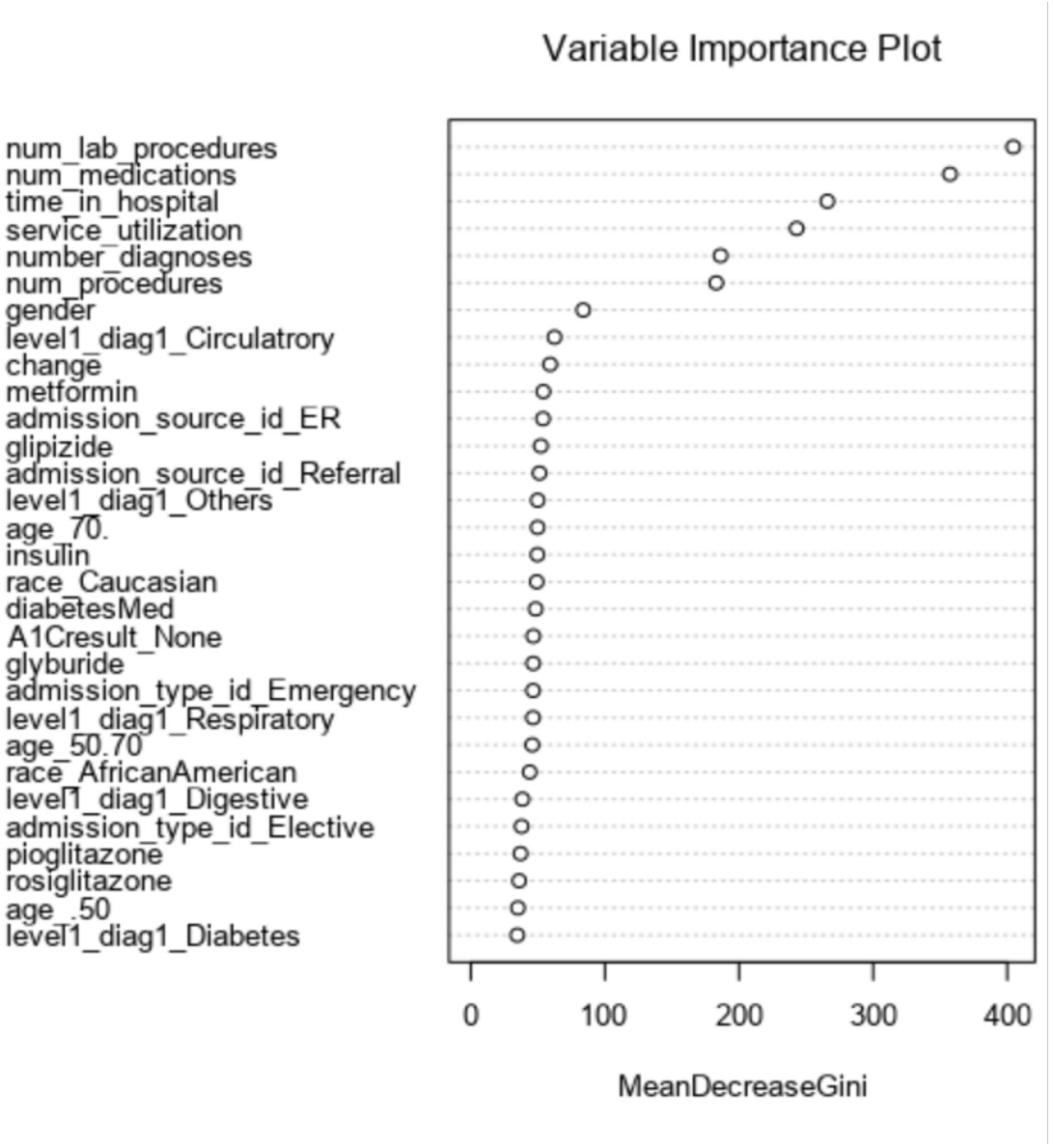
Estimation of error rate: 40.2 %

Confusion Matrix:

	Classification Error	0	1
0	0.442	2259	1790
1	0.363	1481	2603

Model 3 - Random Forest





Model Evaluation

- Since the dataset is unbalanced, accuracy might not be a good measure of accuracy

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_0	Accuracy_1
Logistic_Regression_5	0.6124	0.5938	0.6396	0.6470	0.5766
Decision_Tree_6	0.5912	0.5754	0.6167	0.6180	0.5636
RF1	0.6060	0.6158	0.6438	0.5706	0.6426

Confusion matrix of Decision_Tree_6

	Actual_0	Actual_1
Predicted_0	639	436
Predicted_1	395	563

Confusion matrix of Logistic_Regression_5

	Actual_0	Actual_1
Predicted_0	669	423
Predicted_1	365	576

Confusion matrix of RF1

	Actual_0	Actual_1
Predicted_0	590	357
Predicted_1	444	642

Model Evaluation - Predictions

Actual	
0	1,034
1	999

Logistic_Regression_5	
0	1,092
1	941

Decision_Tree_6	
0	1,075
1	958

RF1	
1	1,086
0	947

Glossary

- RecordID - creates a new column in the data and assigns a unique identifier, that increases sequentially, for each record in the data
- Transpose - pivot the orientation of the data table
- Switch - compares a value against a list of cases and returns the corresponding result