Hospital Readmission Prediction for Diabetic Patients
Report

Candidate Number:HVWM7
Student Number:23229722
COMP0189 - Applied Artificial Intelligence
Coursework 1

I. Task 1: Dataset description

Var.	Miss. Count	Miss. %
WEIGHT	98,569	96.86
MED_SPEC	49,949	49.08
PAY_CODE	40,256	39.56
RACE	3,779	3.71
DIAG_3	1,423	1.40
DIAG_2	358	0.35
DIAG_1	21	0.02
GENDER	3	¡0.01

TABLE I
Missing Data

Variable	Number of Unique Values
DIAG_1	715
DIAG_2	747
DIAG_3	786
MEDICAL_SPECIALTY	72

TABLE II
Unique Values in Variables

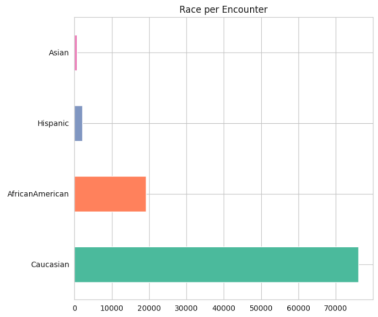

Fig. 1. Distribution of ”Races” - sorted

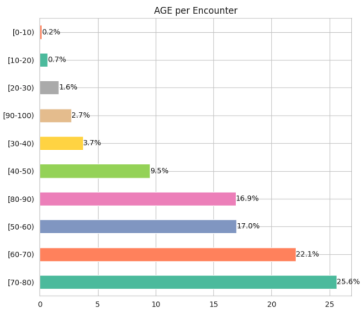

Fig. 2. Distribution of Age - sorted

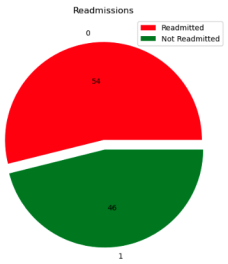

Fig. 3. Pie chart of the Target feature - all encounters

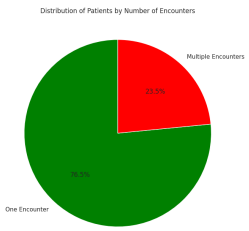

Fig. 4. Pie chart of the Target feature - first encounter

II. Task 2: Data assembling and initial pre-processing

Data Split	Train Shape	Test Shape
TRAINING	(6692, 49)	(2869, 49)
TARGET	(6692,)	(2869,)

TABLE III
TRAINING AND TEST DATA SHAPES

Transformation	Target Col.	Target Value
FeatureDropper	Various	In ‘columns_to_drop‘
SimpleImputer	MED_SPEC	’OTHER’
SimpleImputer	PAYER_CODE	’OTHER’
SimpleImputer	DIAG_1	’OTHER’
SimpleImputer	DIAG_2	’NO DIAG’
SimpleImputer	DIAG_3	’NO DIAG’
SimpleImputer	MAX_GLU	’NO TEST’
SimpleImputer	A1CRES	’NO TEST’
TopCatFilter	MED_SPEC	-
TopCatFilter	PAYER_CODE	-
TopCatFilter	DIAG_1	-
TopCatFilter	DIAG_2	-
TopCatFilter	DIAG_3	-
CustOrdEncoder	Various	In ‘feature_mapping‘

TABLE IV
DATA PREPROCESSING STEPS

III. Task 3 & 4: Design and build a machine learning pipeline
Model Interpretation

	Model	Scaler	accuracy	f1	precision	recall	mean
0	LinearSVC	StandardScaler	0.612550	0.375602	0.531818	0.290323	0.452573
1	LinearSVC	MinMaxScaler	0.612052	0.369231	0.531469	0.282878	0.448907
2	LinearSVC	Normalizer	0.598606	0.000000	0.000000	0.000000	0.149651
3	KNeighborsClassifier	StandardScaler	0.562749	0.405149	0.446269	0.370968	0.446284
4	KNeighborsClassifier	MinMaxScaler	0.552789	0.392422	0.431548	0.359801	0.434140
5	KNeighborsClassifier	Normalizer	0.560259	0.433611	0.448871	0.419355	0.465524
6	SVC	StandardScaler	0.616534	0.368852	0.543478	0.279156	0.452005
7	SVC	MinMaxScaler	0.616036	0.312221	0.555556	0.217122	0.425234
8	SVC	Normalizer	0.598606	0.000000	0.000000	0.000000	0.149651
9	RandomForestClassifier	StandardScaler	0.614044	0.416855	0.529637	0.343672	0.476052
10	RandomForestClassifier	MinMaxScaler	0.614044	0.415976	0.529750	0.342432	0.475550
11	RandomForestClassifier	Normalizer	0.597112	0.083805	0.480519	0.045906	0.301835

Fig. 5. Metrics for four trial models in a non-CV settings, for different scaling methods

Action	**Targe Feature**
Top 10 Categories selection	'MEDICAL_SPECIALTY'
Top 10 Categories selection	'PAYER_CODE'
Top 10 Categories selection	'DIAG_1'
Top 10 Categories selection	'DIAG_2'
Top 10 Categories selection	'DIAG_3'
Ordinal Encoding	'AGE', 'MAX_GLU_SERUM', 'A1CRESULT', 'GLIMEPIRIDE', 'GLIPIZIDE', 'GLYBURIDE', 'INSULIN', 'METFORMIN', 'PIOGLITAZONE', 'REPAGLINIDE', 'ROSIGLITAZONE', 'GENDER', 'CHANGE', 'DIABETESMED'

TABLE V

Preprocessing steps during cross-validation

Classifier	**Parameters**
LinearSVC	model__C: [0.1, 1, 5, 10]
SVC	model__C: [0.1, 1, 5, 10], model__kernel: [linear, rbf]
RandomForestClassifier	model__n_estimators: [10, 50, 100], model__max_depth: [None, 10, 20]
GradientBoostingClassifier	model__n_estimators: [100, 200, 300], model__learning_rate: [0.01, 0.1, 0.2], model__max_depth: [3, 5, 7]

TABLE VI

Parameters for different classifiers

Metric	**Minimum Score** - Model	**Maximum score** - Model
Accuracy	0.605	0.012
f1	0.605	0.011
precision	0.583	0.013
recall **0.610**	0.3 - LinearSVC	0.010

TABLE VII

Minimum and Maximum Test scores for our metrics

A. ***Baseline**: LinearSVC*

Model	Metric	Mean Test Score	Std Test Score	Test Score	model__C
LinearSVC	accuracy	0.606190	0.014726	0.610060	0.100000
LinearSVC	accuracy	0.605977	0.014530	0.610060	0.166810
LinearSVC	accuracy	0.605336	0.014810	0.610060	0.278256
LinearSVC	accuracy	0.605763	0.014813	0.610060	0.464159
LinearSVC	accuracy	0.604909	0.015835	0.610060	0.774264
LinearSVC	accuracy	0.604909	0.015574	0.610060	1.291550
LinearSVC	accuracy	0.600640	0.015964	0.610060	2.154435
LinearSVC	accuracy	0.593170	0.016758	0.610060	3.593814
LinearSVC	accuracy	0.577588	0.008914	0.610060	5.994843
LinearSVC	accuracy	0.553682	0.004744	0.610060	10.000000

Fig. 6. Accuracy scores - LinearSVC

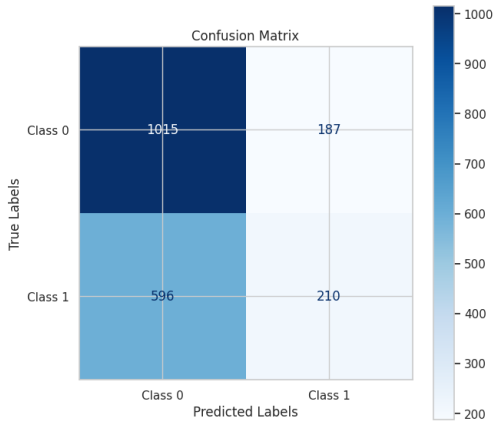

Fig. 7. Confusion Matrix - LinearSVC

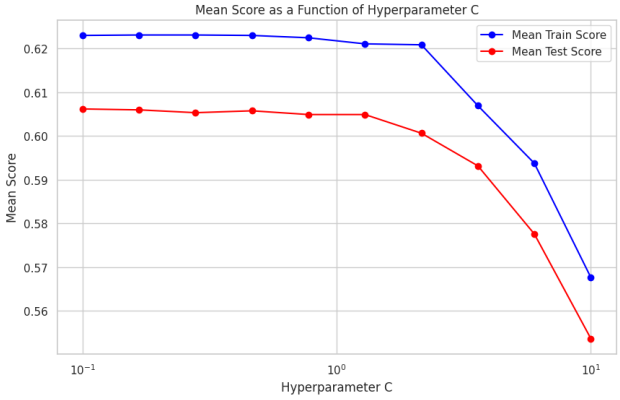

Fig. 8. Performance assessment - LinearSVC

Interpretation

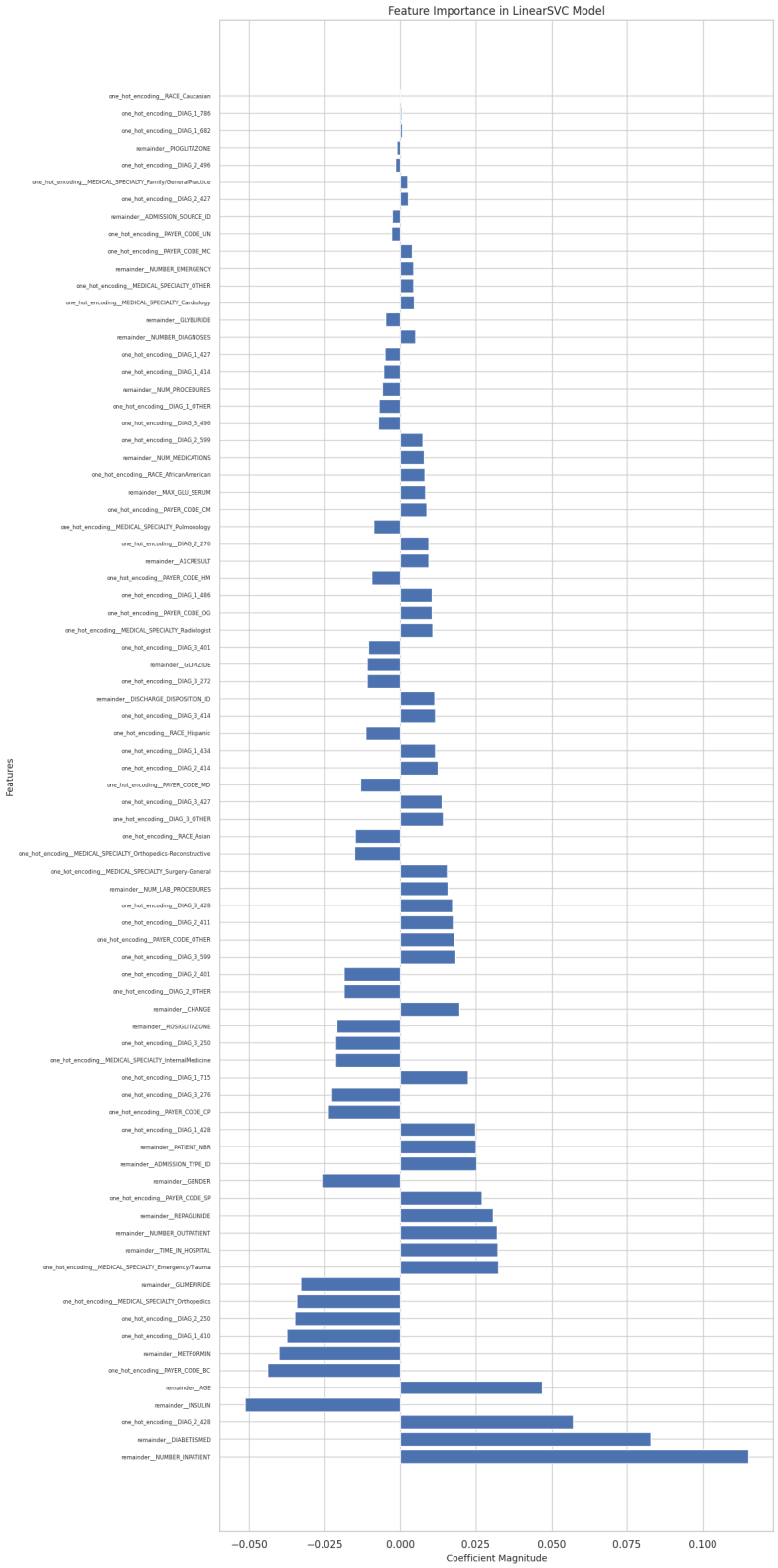

Fig. 9. Coefficient Importance - LinearSVC

B. SVC

Model	Metric	Mean Test Score	Std Test Score	Test Score	C	kernel
SVC	accuracy	0.601921	0.012018	0.612550	0.100000	linear
SVC	accuracy	0.583778	0.000955	0.612550	0.100000	rbf
SVC	accuracy	0.601708	0.011757	0.612550	0.166810	linear
SVC	accuracy	0.588687	0.003734	0.612550	0.166810	rbf
SVC	accuracy	0.600854	0.010821	0.612550	0.278256	linear
SVC	accuracy	0.598719	0.004627	0.612550	0.278256	rbf
SVC	accuracy	0.602348	0.013064	0.612550	0.464159	linear
SVC	accuracy	0.605763	0.002299	0.612550	0.464159	rbf
SVC	accuracy	0.601921	0.012501	0.612550	0.774264	linear
SVC	accuracy	0.603629	0.005123	0.612550	0.774264	rbf
SVC	accuracy	0.603415	0.009888	0.612550	1.291550	linear
SVC	accuracy	0.607257	0.006544	0.612550	1.291550	rbf
SVC	accuracy	0.603415	0.013064	0.612550	2.154435	linear
SVC	accuracy	0.601067	0.012508	0.612550	2.154435	rbf
SVC	accuracy	0.601921	0.012018	0.612550	3.593814	linear
SVC	accuracy	0.593170	0.013757	0.612550	3.593814	rbf
SVC	accuracy	0.602134	0.011815	0.612550	5.994843	linear
SVC	accuracy	0.580363	0.019168	0.612550	5.994843	rbf
SVC	accuracy	0.602988	0.012948	0.612550	10.000000	linear
SVC	accuracy	0.573746	0.016404	0.612550	10.000000	rbf

Fig. 10. Accuracy scores - SVC

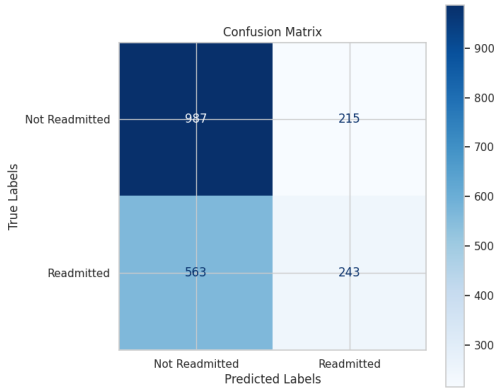

Fig. 11. Confusion Matrix - SVC

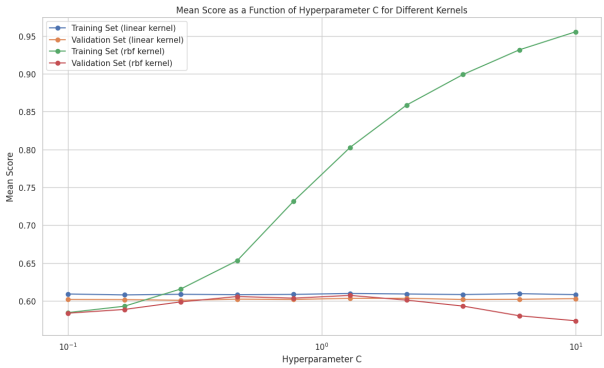

Fig. 12. Performance assessment - SVC

Interpretation

No Coefficient Importance as the best model was not linear not Tree-based

C. RandomForestClassifier

	Model	Metric	Mean Test Score	Std Test Score	Test Score	max_depth	n_estimators
	RandomForestClassifier	accuracy	0.579723	0.009998	0.620518	nan	10
	RandomForestClassifier	accuracy	0.601281	0.010513	0.620518	nan	30
	RandomForestClassifier	accuracy	0.611099	0.015838	0.620518	nan	50
	RandomForestClassifier	accuracy	0.613874	0.016987	0.620518	nan	70
	RandomForestClassifier	accuracy	0.619637	0.012530	0.620518	nan	100
	RandomForestClassifier	accuracy	0.620064	0.016847	0.620518	nan	150
	RandomForestClassifier	accuracy	0.605977	0.009363	0.620518	5.000000	10
	RandomForestClassifier	accuracy	0.602775	0.005938	0.620518	5.000000	30
	RandomForestClassifier	accuracy	0.603415	0.004508	0.620518	5.000000	50
	RandomForestClassifier	accuracy	0.602134	0.006962	0.620518	5.000000	70
	RandomForestClassifier	accuracy	0.599787	0.005923	0.620518	5.000000	100
	RandomForestClassifier	accuracy	0.599787	0.005272	0.620518	5.000000	150
	RandomForestClassifier	accuracy	0.595731	0.010787	0.620518	10.000000	10
	RandomForestClassifier	accuracy	0.604269	0.005783	0.620518	10.000000	30
	RandomForestClassifier	accuracy	0.602561	0.004754	0.620518	10.000000	50
	RandomForestClassifier	accuracy	0.605336	0.005459	0.620518	10.000000	70
	RandomForestClassifier	accuracy	0.603202	0.004128	0.620518	10.000000	100
	RandomForestClassifier	accuracy	0.604909	0.008167	0.620518	10.000000	150
	RandomForestClassifier	accuracy	0.587407	0.006460	0.620518	15.000000	10
	RandomForestClassifier	accuracy	0.605550	0.010791	0.620518	15.000000	30
	RandomForestClassifier	accuracy	0.616649	0.006201	0.620518	15.000000	50
	RandomForestClassifier	accuracy	0.621985	0.010934	0.620518	15.000000	70
	RandomForestClassifier	accuracy	0.622199	0.012224	0.620518	15.000000	100
	RandomForestClassifier	accuracy	0.620491	0.010854	0.620518	15.000000	150

Fig. 13. Accuracy scores - Random Forest

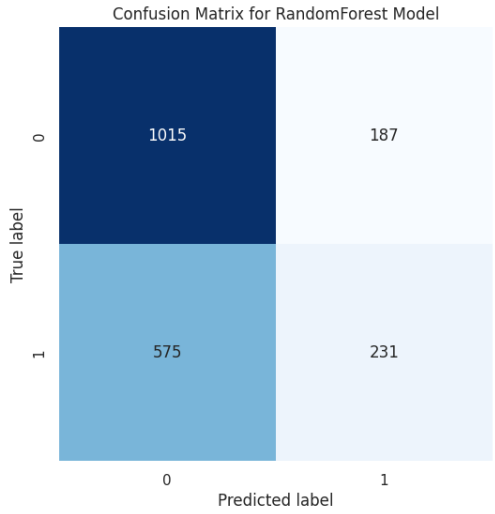

Fig. 14. Confusion Matrix - Random Forest

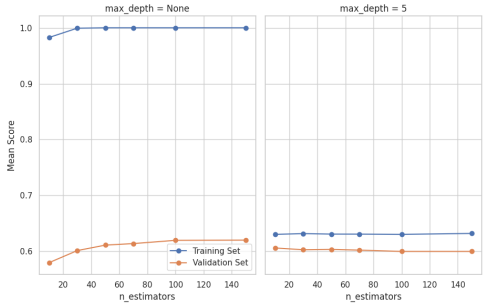

Fig. 15. Performance Assessment - max_depth=['None',5] - Random Forest

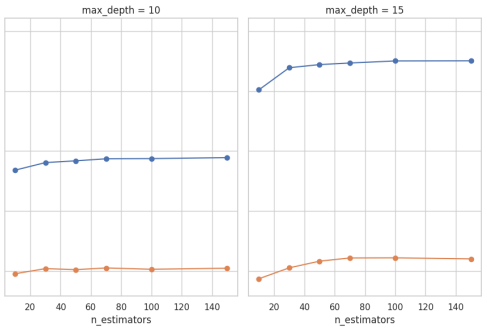

Fig. 16. Performance Assessment - max_depth=[10,15] - Random Forest

Interpretation

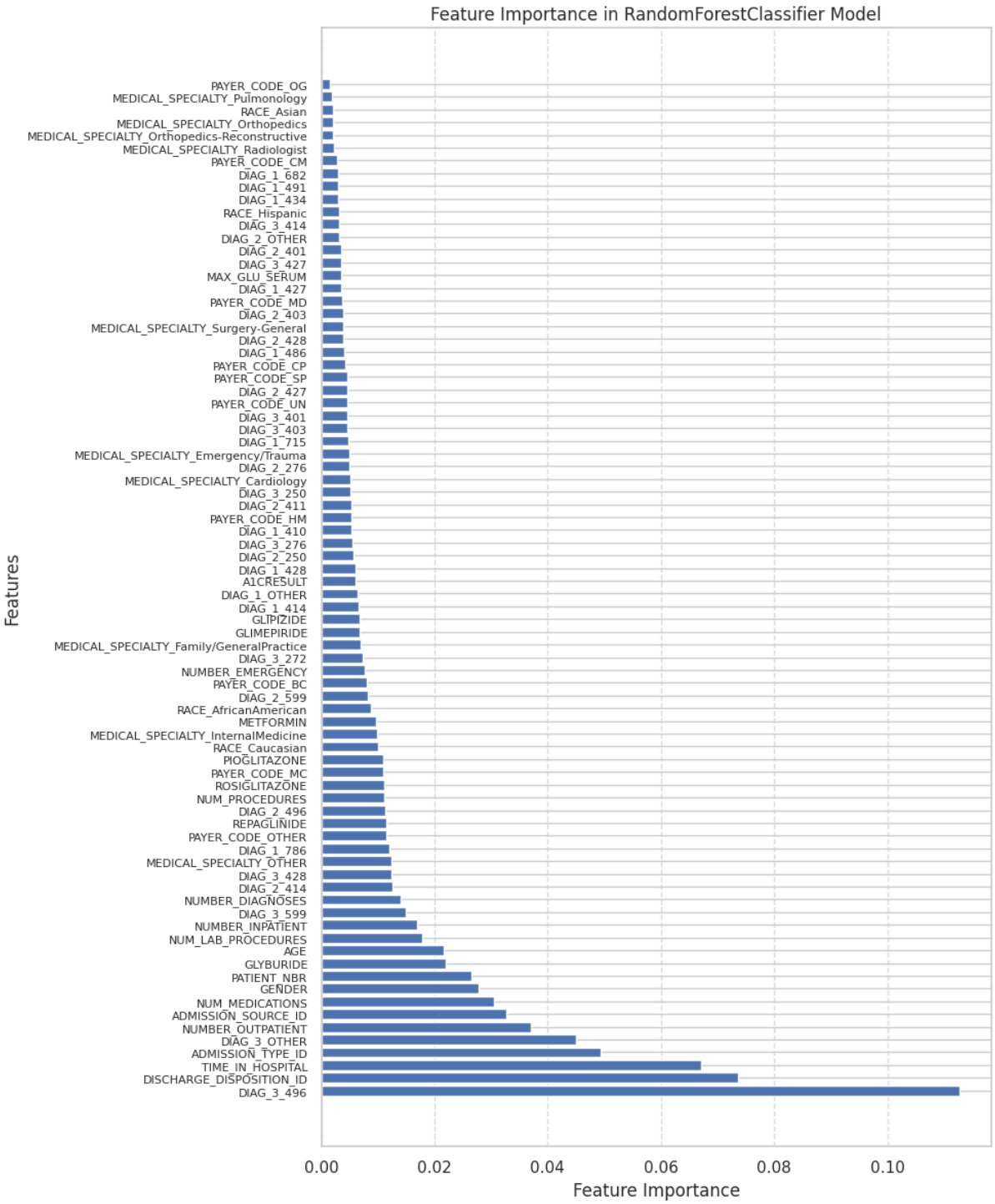

Fig. 17. Feature Importance - Random Forest

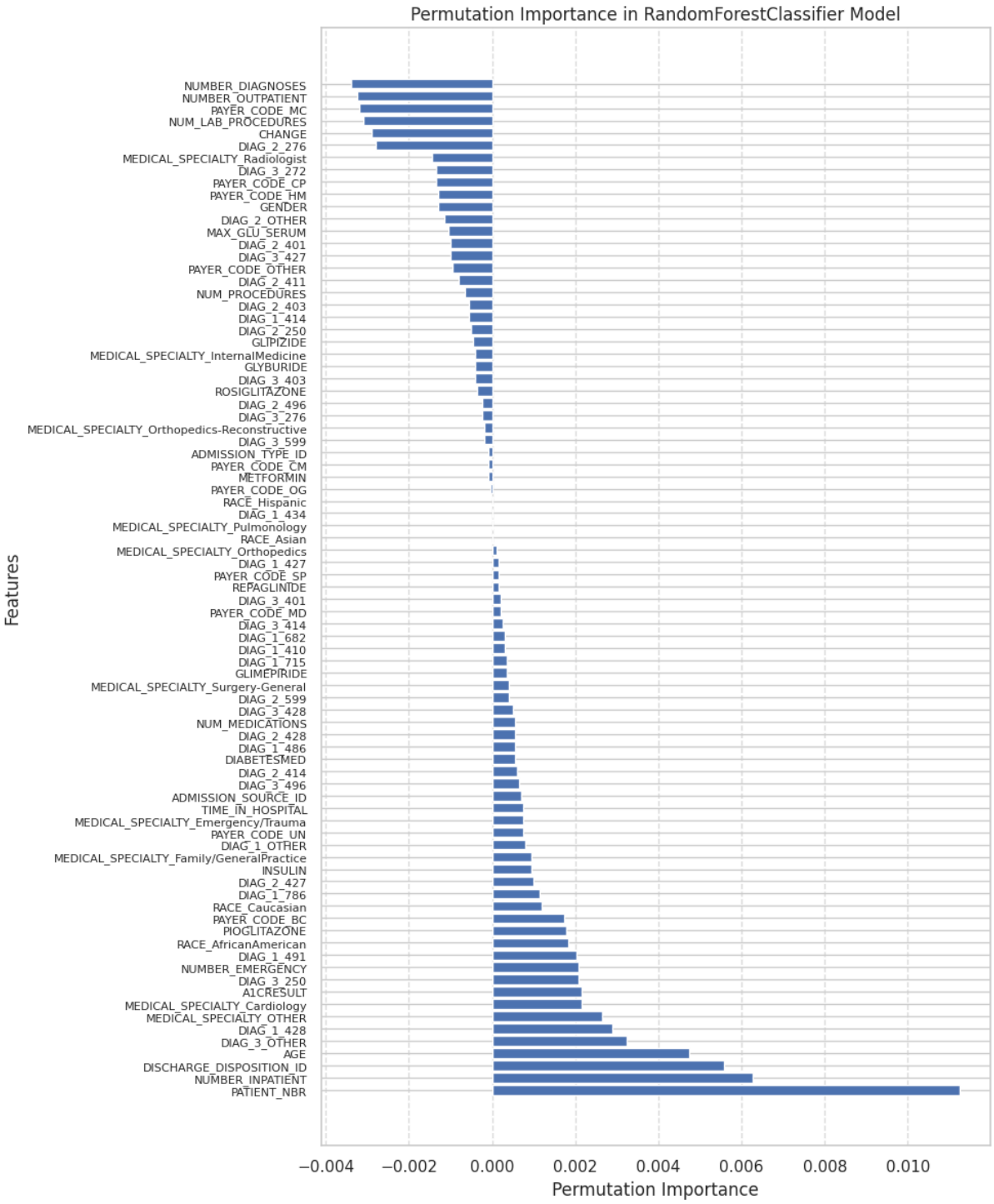

Fig. 18. Permutation Feature Importance - Random Forest

D. *GradientBoosting*

	Model	Metric	Mean Test Score	Std Test Score	Test Score	learning_rate	max_depth	n_estimators
	GradientBoostingClassifier	accuracy	0.605123	0.007900	0.611056	0.010000	3	100
	GradientBoostingClassifier	accuracy	0.607684	0.007131	0.611056	0.010000	3	150
	GradientBoostingClassifier	accuracy	0.610672	0.009240	0.611056	0.010000	3	200
	GradientBoostingClassifier	accuracy	0.612166	0.012906	0.611056	0.010000	3	250
	GradientBoostingClassifier	accuracy	0.613447	0.013147	0.611056	0.010000	3	300
	GradientBoostingClassifier	accuracy	0.604482	0.009171	0.611056	0.010000	4	100
	GradientBoostingClassifier	accuracy	0.609819	0.008256	0.611056	0.010000	4	150
	GradientBoostingClassifier	accuracy	0.613661	0.012206	0.611056	0.010000	4	200
	GradientBoostingClassifier	accuracy	0.615368	0.013510	0.611056	0.010000	4	250
	GradientBoostingClassifier	accuracy	0.617503	0.009631	0.611056	0.010000	4	300
	GradientBoostingClassifier	accuracy	0.604696	0.007918	0.611056	0.010000	5	100
	GradientBoostingClassifier	accuracy	0.612593	0.008955	0.611056	0.010000	5	150
	GradientBoostingClassifier	accuracy	0.616009	0.011423	0.611056	0.010000	5	200
	GradientBoostingClassifier	accuracy	0.615368	0.012817	0.611056	0.010000	5	250
	GradientBoostingClassifier	accuracy	0.616222	0.017388	0.611056	0.010000	5	300
	GradientBoostingClassifier	accuracy	0.604909	0.006930	0.611056	0.010000	6	100
	GradientBoostingClassifier	accuracy	0.613020	0.011083	0.611056	0.010000	6	150
	GradientBoostingClassifier	accuracy	0.615795	0.012280	0.611056	0.010000	6	200
	GradientBoostingClassifier	accuracy	0.610672	0.011718	0.611056	0.010000	6	250
	GradientBoostingClassifier	accuracy	0.612380	0.019036	0.611056	0.010000	6	300
	GradientBoostingClassifier	accuracy	0.601494	0.007289	0.611056	0.010000	7	100
	GradientBoostingClassifier	accuracy	0.607471	0.013337	0.611056	0.010000	7	150
	GradientBoostingClassifier	accuracy	0.610459	0.009943	0.611056	0.010000	7	200
	GradientBoostingClassifier	accuracy	0.611740	0.013043	0.611056	0.010000	7	250
	GradientBoostingClassifier	accuracy	0.610245	0.011120	0.611056	0.010000	7	300
	GradientBoostingClassifier	accuracy	0.617503	0.014565	0.611056	0.050000	3	100
	GradientBoostingClassifier	accuracy	0.619424	0.014184	0.611056	0.050000	3	150
	GradientBoostingClassifier	accuracy	0.619637	0.008447	0.611056	0.050000	3	200
	GradientBoostingClassifier	accuracy	0.621345	0.010245	0.611056	0.050000	3	250

Fig. 19. Accuracy scores - Gradient Boosting

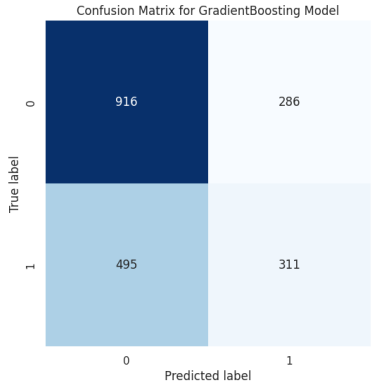

Fig. 20. Confusion Matrix - Gradient Boosting

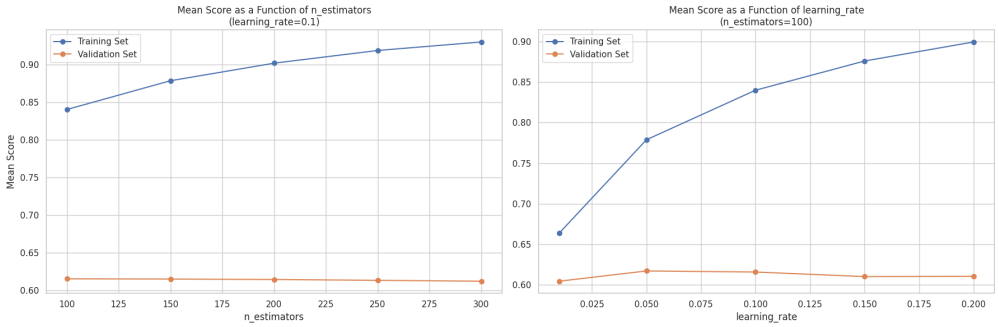

Fig. 21. Performances as a function of n_estimator and learning_rate - GradientBoosting

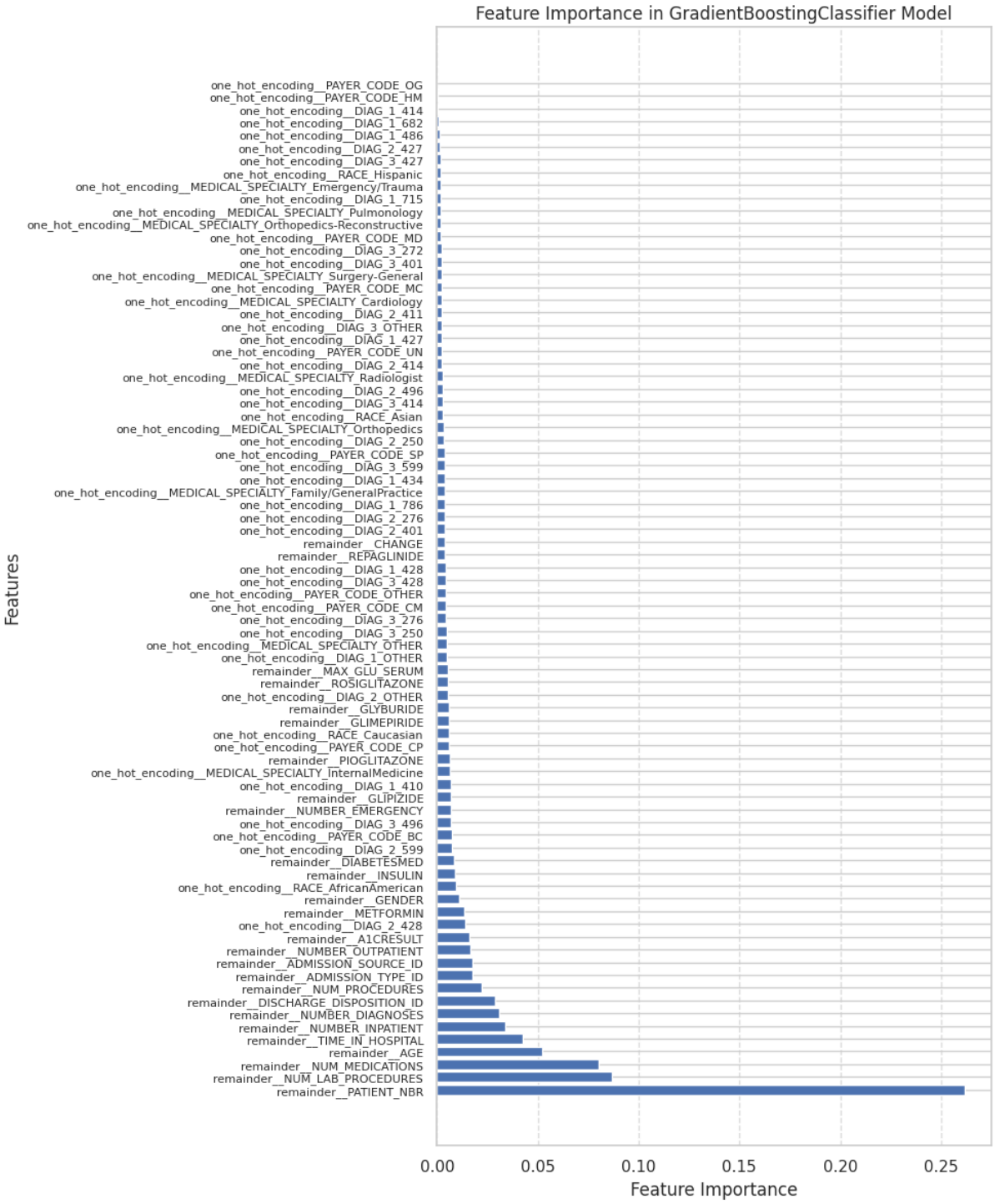

Fig. 22. Feature Importance - Gradient Boosting

IV. Identify limitations and propose potential solutions

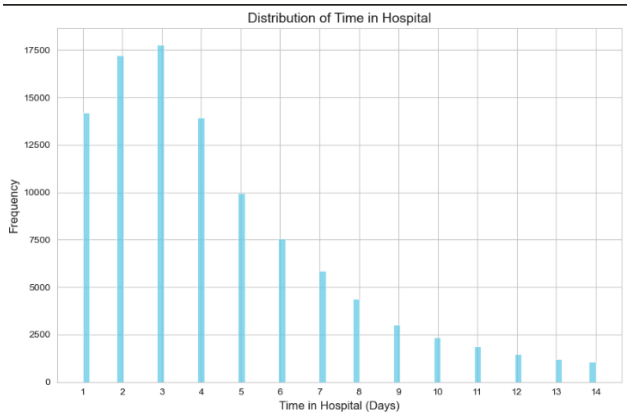

Fig. 23. Distribution of the time spent in hospital

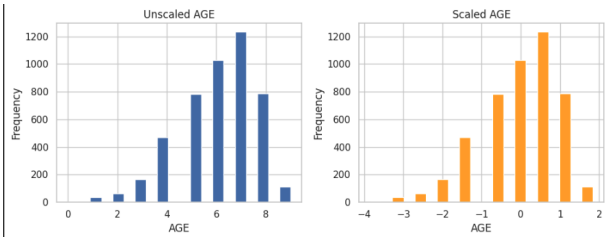

Fig. 24. Scaling example