PREDICTING HOSPITAL READMISSION FOR PATIENTS WITH DIABETICS

A Project report submitted to

DEPARTMENT OF

COMPUTER SCIENCE AND ENGINEERING

In partial fulfillment of the requirements for the award of the degree of

Bachelor Of Technology In

Computer Science And Engineering

CHALAPATHI INSTITUTE OF ENGINEERING & TECHNOLOGY

Presented by

CH.RENUKA	**Y16CS1219**
B.MOUNIKA	**Y16CS1212**
M.UNNATHA	**Y16CS1263**
B.GOPI KRISHNA	**Y16CS1216**

Under the esteemed guidance of

Dr.B.V.V.S.PRASAD,

Associate professor,

Dept. of CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CHALAPATHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

(Approved by A.I.C.T.E, Affiliated To Acharya Nagarjuna University)

GUNTUR-522034

2019-2020

Introduction

As the healthcare system moves toward value-based care, CMS has created many programs to improve the quality of care of patients. One of these programs is called the Hospital Readmission Reduction Program (HRRP), which reduces reimbursement to hospitals with above average re admissions. For those hospitals which are currently penalized under this program, one solution is to create interventions to provide additional assistance to patients with increased risk of readmission. But how do we identify these patients? We can use predictive modeling from data science to help prioritize patients.

One patient population that is at increased risk of hospitalization and readmission is that of diabetes. Diabetes is a medical condition that affects approximately 1 in 10 patients in the United States. According to Ostling et al, patients with diabetes have almost double the chance of being hospitalized than the general population (Ostling et al 2017). Therefore, in this article, I will focus on predicting hospital readmission for patients with diabetes.

In this project I will demonstrate how to build a model predicting readmission in Python using the following steps

- data exploration
- feature engineering
- building training/validation/test samples
- model selection
- model evaluation

You can follow along with the Jupyter Notebook provided on my github (https://github.com/andrewwlong/diabetes_readmission).

Data Exploration

The data that is used in this project originally comes from the UCI machine learning repository

Feature Engineering

In this section, we will create features for our predictive model. For each section, we will add new variables to the dataframe and then keep track of which columns of the dataframe we want to use as part of the predictive model features. We will break down this section into numerical features, categorical features and extra features.

Numerical Features

The easiest type of features to use is numerical features. These features do not need any modification.

Categorical Features

The next type of features we want to create are categorical variables. Categorical variables are non-numeric data such as race and gender. To turn these non-numerical data into variables, the simplest thing is to use a technique called one-hot encoding.

Extra Features

The last two columns we want to make features are `age` and `weight`. Typically, you would think of these as numerical data.

Model Selection: Baseline models

In this section, we will first compare the performance of the following 7 machine learning models using default hyperparameters:

- K-nearest neighbors
- Logistic regression
- Stochastic gradient descent
- Naive Bayes
- Decision tree
- Random forest
- Gradient boosting classifier

Conclusion

Through this project, we created a machine learning model that is able to predict the patients with diabetes with highest risk of being readmitted within 30 days. The best model was a gradient boosting classifier with optimized hyperparameters. The model was able to catch 58% of the readmissions and is about 1.5 times better than just randomly picking patients. Overall, I believe many healthcare data scientists are working on predictive models for hospital readmission.

screen shots of outputs:

In [5]:

```
# count the number of rows for each type
df.groupby('readmitted').size()
```

Out[5]:

```
readmitted
<30    11357
>30    35545
NO     54864
dtype: int64
```

In [24]:

```
df[cols_cat].isnull().sum()
```

Out[24]:

```
race                2234
```

```
gender                      0
max_glu_serum               0
A1Cresult                   0
metformin                   0
repaglinide                 0
nateglinide                 0
chlorpropamide              0
glimepiride                 0
acetohexamide               0
glipizide                   0
glyburide                   0
tolbutamide                 0
pioglitazone                0
rosiglitazone               0
acarbose                    0
miglitol                    0
troglitazone                0
tolazamide                  0
insulin                     0
glyburide-metformin         0
glipizide-metformin         0
glimepiride-pioglitazone    0
metformin-rosiglitazone     0
metformin-pioglitazone      0
change                      0
diabetesMed                 0
payer_code              39398
dtype: int64
```

In [27]:

```

```

In [20]:

```
df.groupby('med_spec').size()
```

Out[20]:

```
med_spec
Cardiology                  5279
Emergency/Trauma            7419
Family/GeneralPractice      7252
InternalMedicine           14237
Nephrology                  1539
```

```
Orthopedics                 1392
Orthopedics-Reconstructive     1230
Other                    8199
Radiologist                 1121
Surgery-General              3059
UNK                     48616
dtype: int64
```

```
feature_importances.head()
```

Out[74]:

	importance
number_inpatient	0.356977
rosiglitazone_No	0.283933
rosiglitazone_Steady	0.238041
discharge_disposition_id_22	0.202501
repaglinide_No	0.170694

```
rf.get_params()
```

Out[80]:

```
{'bootstrap': True,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': 6,
 'max_features': 'auto',
 'max_leaf_nodes': None,
 'min_impurity_decrease': 0.0,
 'min_impurity_split': None,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 10,
 'n_jobs': 1,
 'oob_score': False,
 'random_state': 42,
 'verbose': 0,
 'warm_start': False}
```

In [88]:

```
sgdc_random.best_params_
```

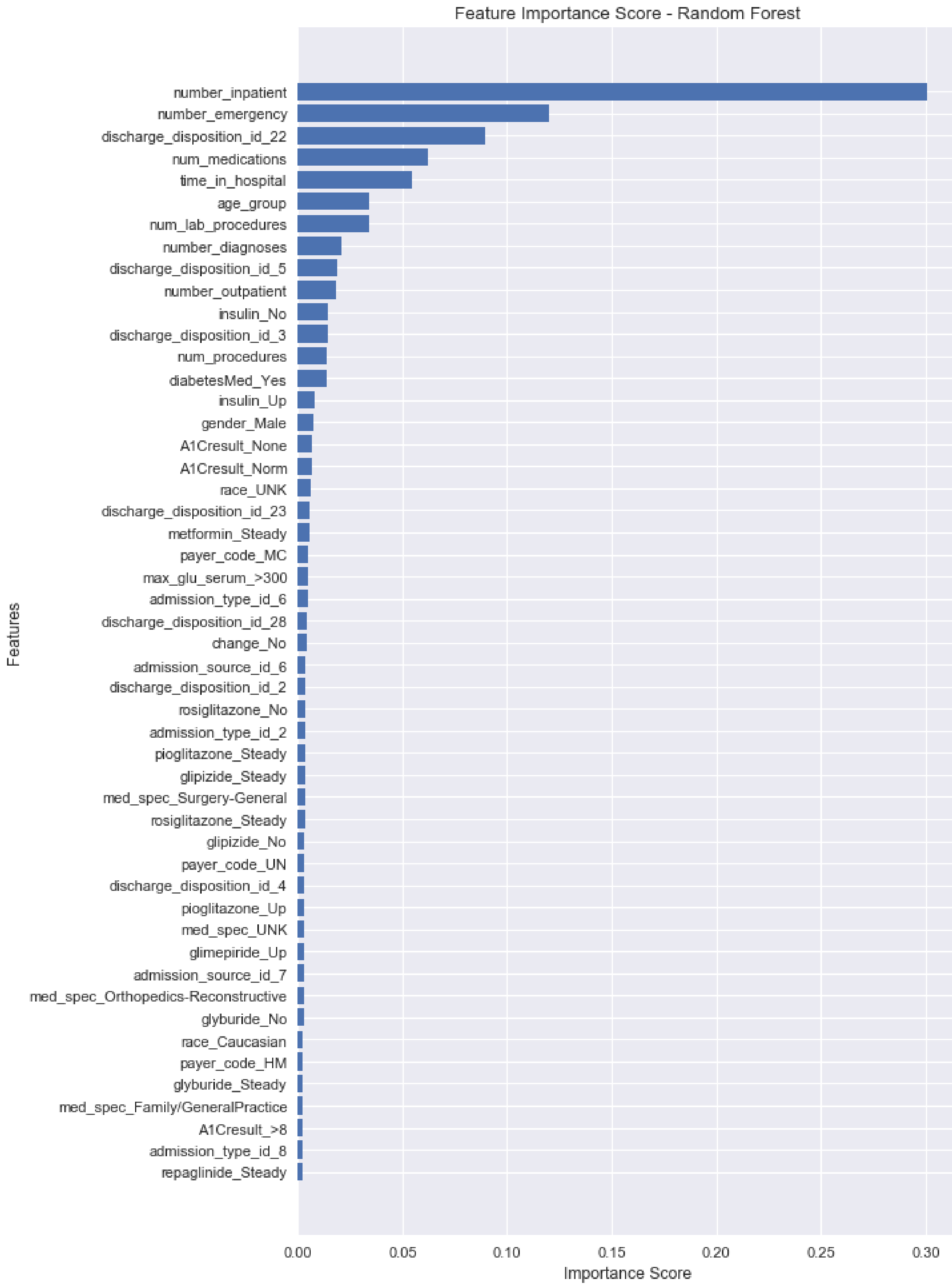
Out[88]:

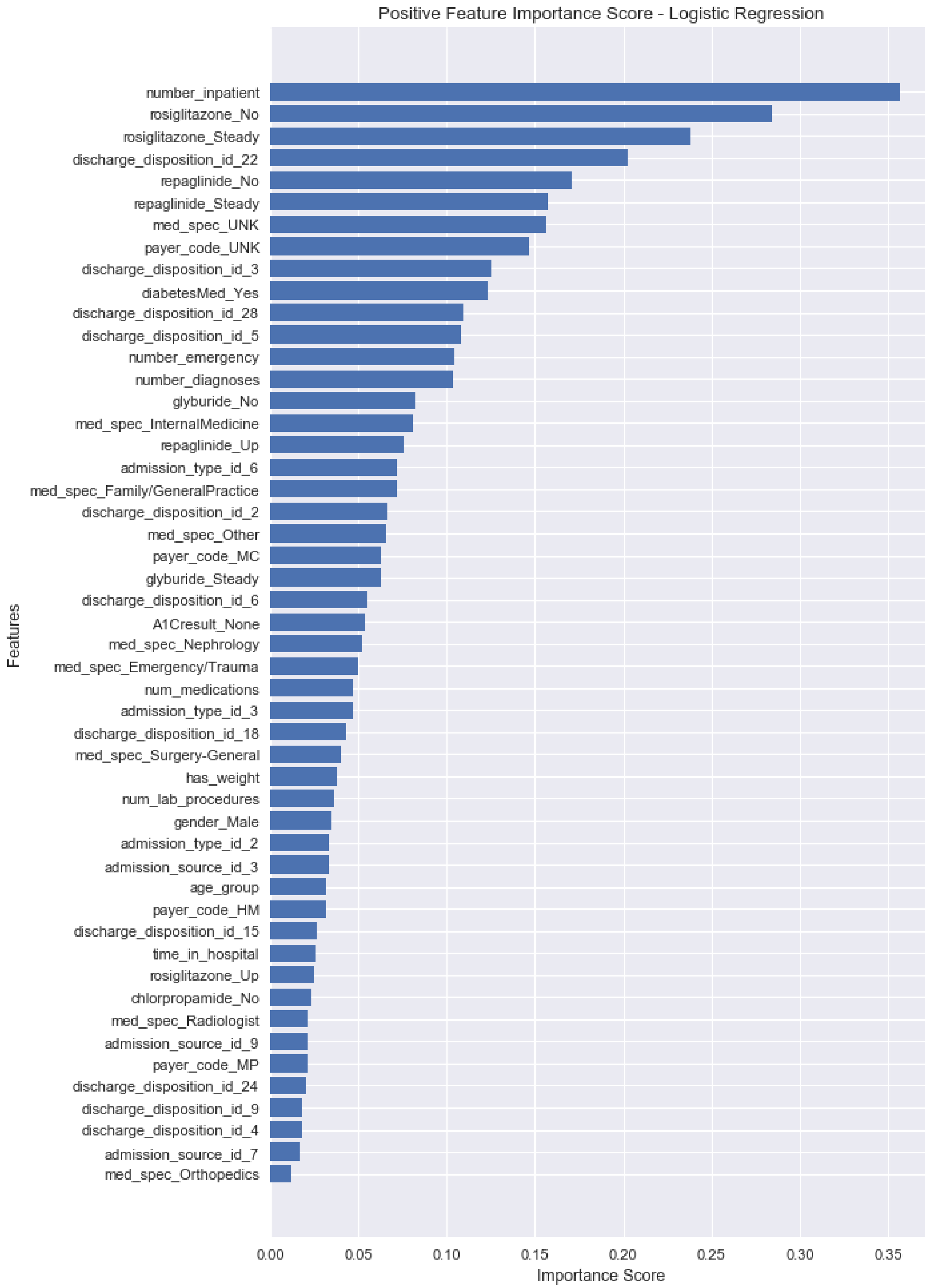
{'alpha': 0.001, 'max_iter': 200, 'penalty': 'l1'}

```
df_results
```

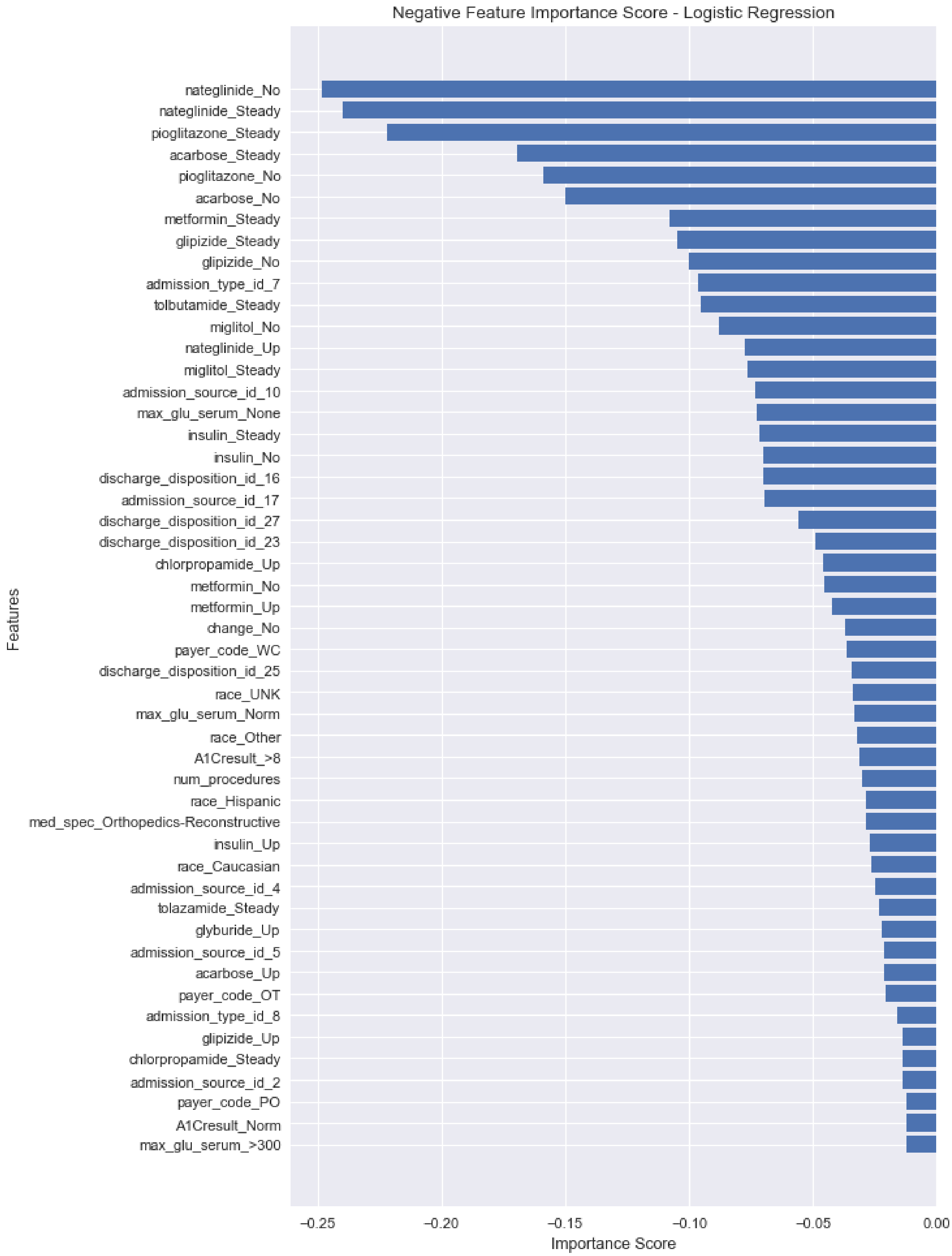
Out[94]:

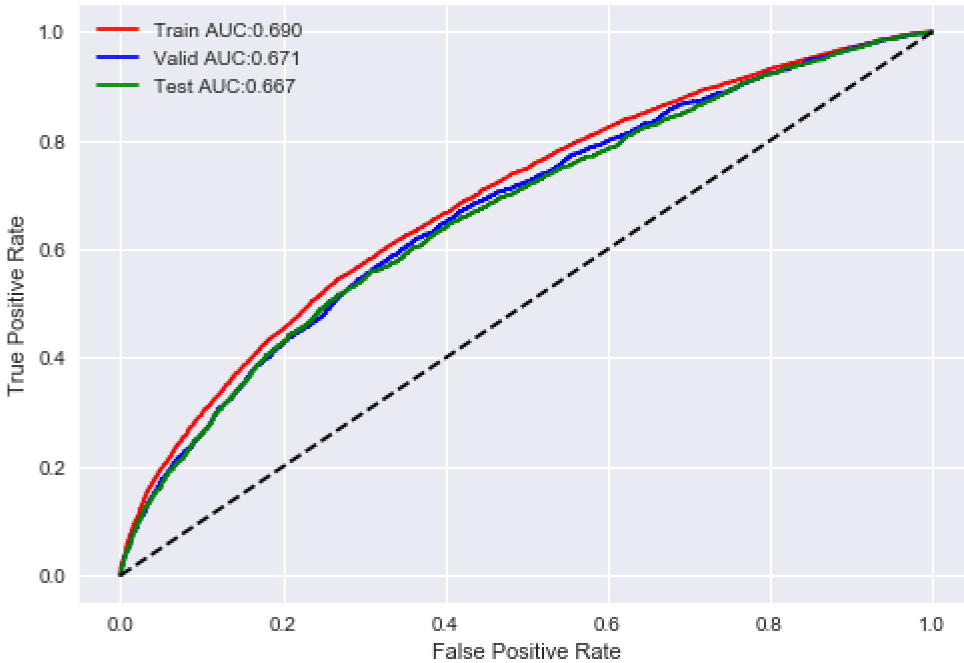
	auc	classifier	data_set
0	0.661886	SGD	base
1	0.663974	SGD	optimized
2	0.648315	RF	base
3	0.660519	RF	optimized
4	0.638967	GB	base
5	0.671137	GB	optimized


Feature Importance Score - Random Forest
Features
Importance Score
number_inpatient
number_emergency
discharge_disposition_id_22
num_medications
time_in_hospital
age_group
num_lab_procedures
number_diagnoses
discharge_disposition_id_5
number_outpatient
insulin_No
discharge_disposition_id_3
num_procedures
diabetesMed_Yes
insulin_Up
gender_Male
A1Cresult_None
A1Cresult_Norm
race_UNK
discharge_disposition_id_23
metformin_Steady
payer_code_MC
max_glu_serum_>300
admission_type_id_6
discharge_disposition_id_28
change_No
admission_source_id_6
discharge_disposition_id_2
rosiglitazone_No
admission_type_id_2
pioglitazone_Steady
glipizide_Steady
med_spec_Surgery-General
rosiglitazone_Steady
glipizide_No
payer_code_UN
discharge_disposition_id_4
pioglitazone_Up
med_spec_UNK
glimepiride_Up
admission_source_id_7
med_spec_Orthopedics-Reconstructive
glyburide_No
race_Caucasian
payer_code_HM
glyburide_Steady
med_spec_Family/GeneralPractice
A1Cresult_>8
admission_type_id_8
repaglinide_Steady
0.00
0.05
0.10
0.15
0.20
0.25
0.30



Positive Feature Importance Score - Logistic Regression
number_inpatient
rosiglitazone_No
rosiglitazone_Steady
discharge_disposition_id_22
repaglinide_No
repaglinide_Steady
med_spec_UNK
payer_code_UNK
discharge_disposition_id_3
diabetesMed_Yes
discharge_disposition_id_28
discharge_disposition_id_5
number_emergency
number_diagnoses
glyburide_No
med_spec_InternalMedicine
repaglinide_Up
admission_type_id_6
med_spec_Family/GeneralPractice
discharge_disposition_id_2
med_spec_Other
payer_code_MC
glyburide_Steady
discharge_disposition_id_6
A1Cresult_None
med_spec_Nephrology
med_spec_Emergency/Trauma
num_medications
admission_type_id_3
discharge_disposition_id_18
med_spec_Surgery-General
has_weight
num_lab_procedures
gender_Male
admission_type_id_2
admission_source_id_3
age_group
payer_code_HM
discharge_disposition_id_15
time_in_hospital
rosiglitazone_Up
chlorpropamide_No
med_spec_Radiologist
admission_source_id_9
payer_code_MP
discharge_disposition_id_24
discharge_disposition_id_9
discharge_disposition_id_4
admission_source_id_7
med_spec_Orthopedics
Features
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Importance Score


Negative Feature Importance Score - Logistic Regression
Features
nateglinide_No
nateglinide_Steady
pioglitazone_Steady
acarbose_Steady
pioglitazone_No
acarbose_No
metformin_Steady
glipizide_Steady
glipizide_No
admission_type_id_7
tolbutamide_Steady
miglitol_No
nateglinide_Up
miglitol_Steady
admission_source_id_10
max_glu_serum_None
insulin_Steady
insulin_No
discharge_disposition_id_16
admission_source_id_17
discharge_disposition_id_27
discharge_disposition_id_23
chlorpropamide_Up
metformin_No
metformin_Up
change_No
payer_code_WC
discharge_disposition_id_25
race_UNK
max_glu_serum_Norm
race_Other
A1Cresult_>8
num_procedures
race_Hispanic
med_spec_Orthopedics-Reconstructive
insulin_Up
race_Caucasian
admission_source_id_4
tolazamide_Steady
glyburide_Up
admission_source_id_5
acarbose_Up
payer_code_OT
admission_type_id_8
glipizide_Up
chlorpropamide_Steady
admission_source_id_2
payer_code_PO
A1Cresult_Norm
max_glu_serum_>300
−0.25
−0.20
−0.15
−0.10
−0.05
0.00
Importance Score



Train AUC:0.690
Valid AUC:0.671
Test AUC:0.667
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0