Predicting hospital readmission within 30 days of discharge for diabetic patients

Ignasi Sols

Introduction

9.3%

U.S population that have Diabetes Mellitus

+

~14-22%

30-day readmission rates for hospitalized patients with diabetes

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

Medicare

Does not reimburse hospitals for readmissions that happen within 30 days

UPMC LIFE CHANGING MEDICINE

Business impact: Reduce the number of healthcare claims for diabetic patients, and increase the percentage of Medicare reimbursements to hospitals

Impact hypothesis:

Providing specialized diabetic and lifestyle education to patients at higher risk of hospital readmission will reduce the readmission rates.

Data Science Goal:

Predict hospital readmission within 30 days of discharge

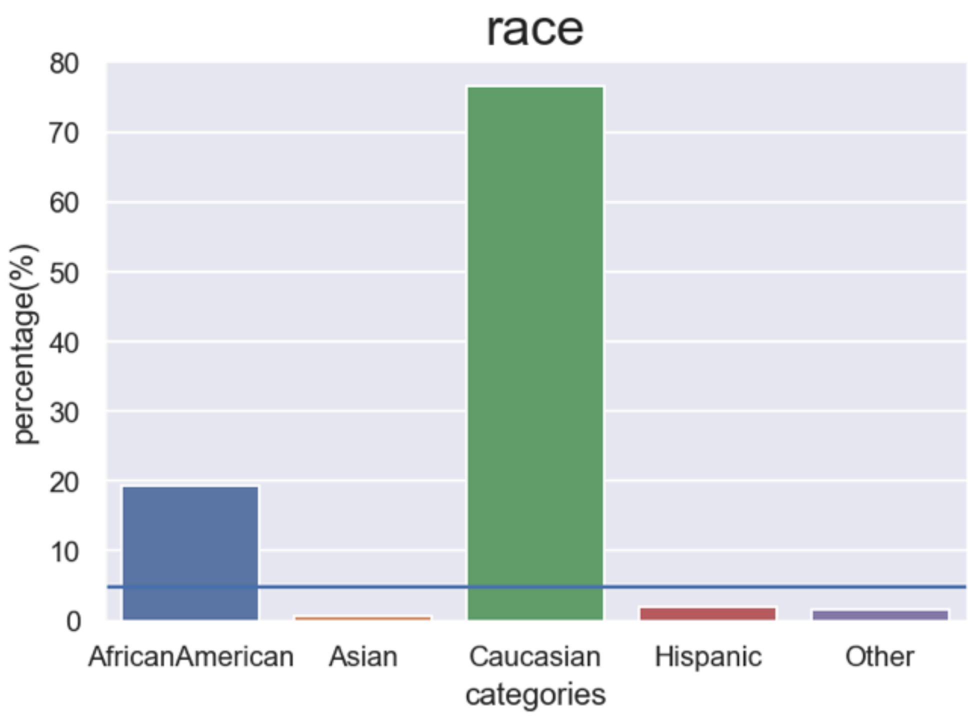
Dataset: UC-Irvine/Kaggle (Diabetes 130-US hospitals for years 1999-2008)

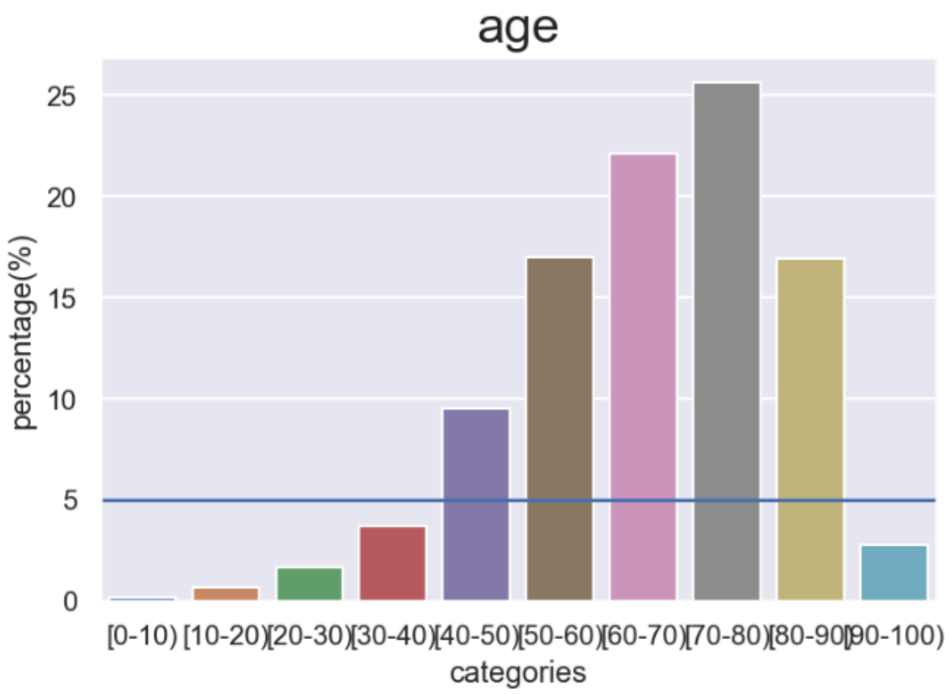
~100K hospital patient admissions (rows). Some patients were admitted more than once (they appear in more than one row).

Tools: Pandas, sklearn

Assumptions/Risks

-Different historical context (the dataset, from 1999 to 2008, might be outdated as science and medicine advance fast.
-Data from these 130 hospitals might not generalize to the UMPC hospital.
-Most data was from caucasian patients. The results might not generalize well to other races and patients younger than 50 years old



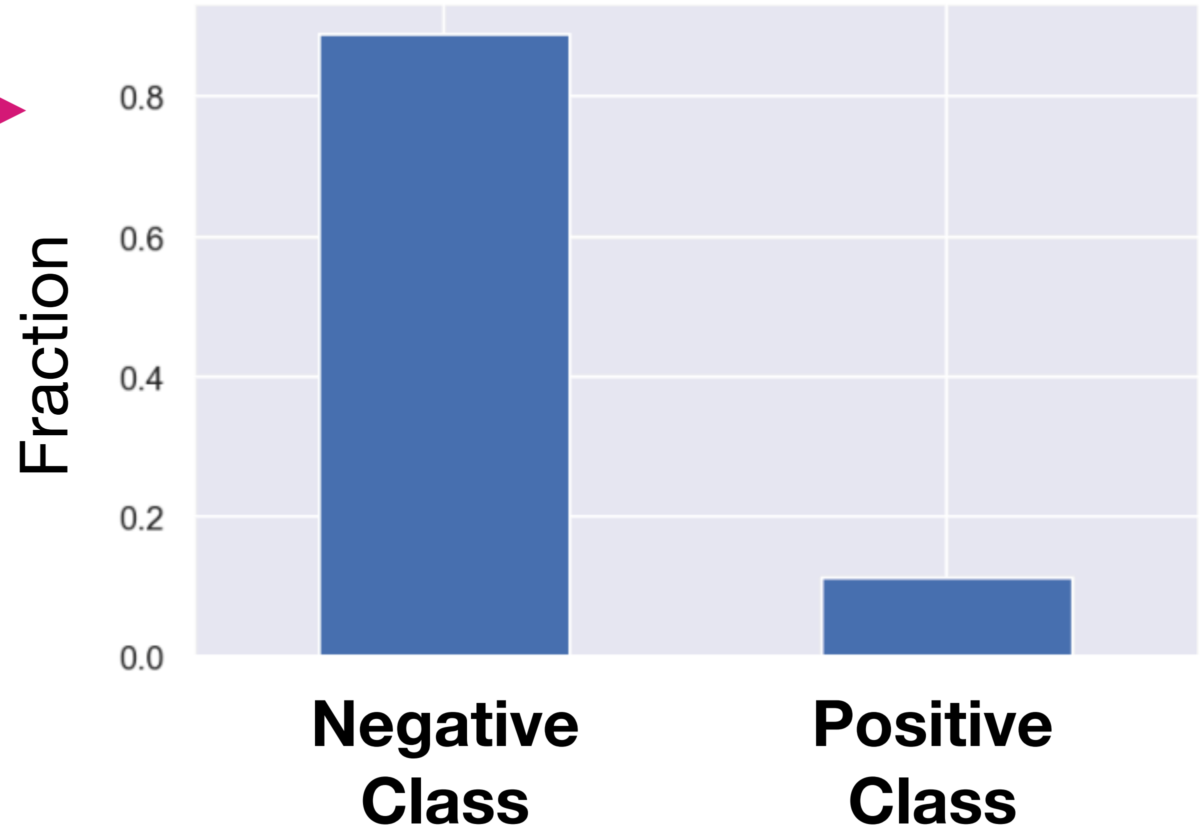


Data Science path:

Build a machine learning classification model that predicts hospital readmission.

-The target had three classes: I binarized it:
Positive class: <30 days
Negative class: >30 days & ‘no record of readmission’.

-Highly **imbalanced** data. This makes the analyses more challenging.



-50 features, mostly categorical

-Trained *Logistic Regression* and *Random forest* models

UPMC LIFE CHANGING MEDICINE

For each ~100 patients that will be readmitted within 30 days Of discharge, our model will detect ~47 readmissions.

Assuming a 50% success of the education program

↓

~ 24 + claims that are reimbursed monthly
Huge savings!

If specialized diabetic and lifestyle education program takes ~30 min/patient

↓ ↓

Naive model (giving education to all diabetic patients) ~467h/month

Our model ~138.5h/ month

*Classification metric: **F1 score = 0.25**

The final model had 3 features:

Number of inpatient visits of the patient in the year preceding the encounter

Number of emergency visits of the patient in the year preceding the encounter

Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values

Future directions:

-The target variable has three classes, that I binarized. Should I treat it as a multi-class problem?
-Optimize with linear regression as well, instead of just random forests.
-Run XGBoost with logistic regression, ensambling with different models.
-Deal differently with categorical variables when running Random Forests (use LabelEncoder instead of making dummy variables?)

Thanks!

Appendix:

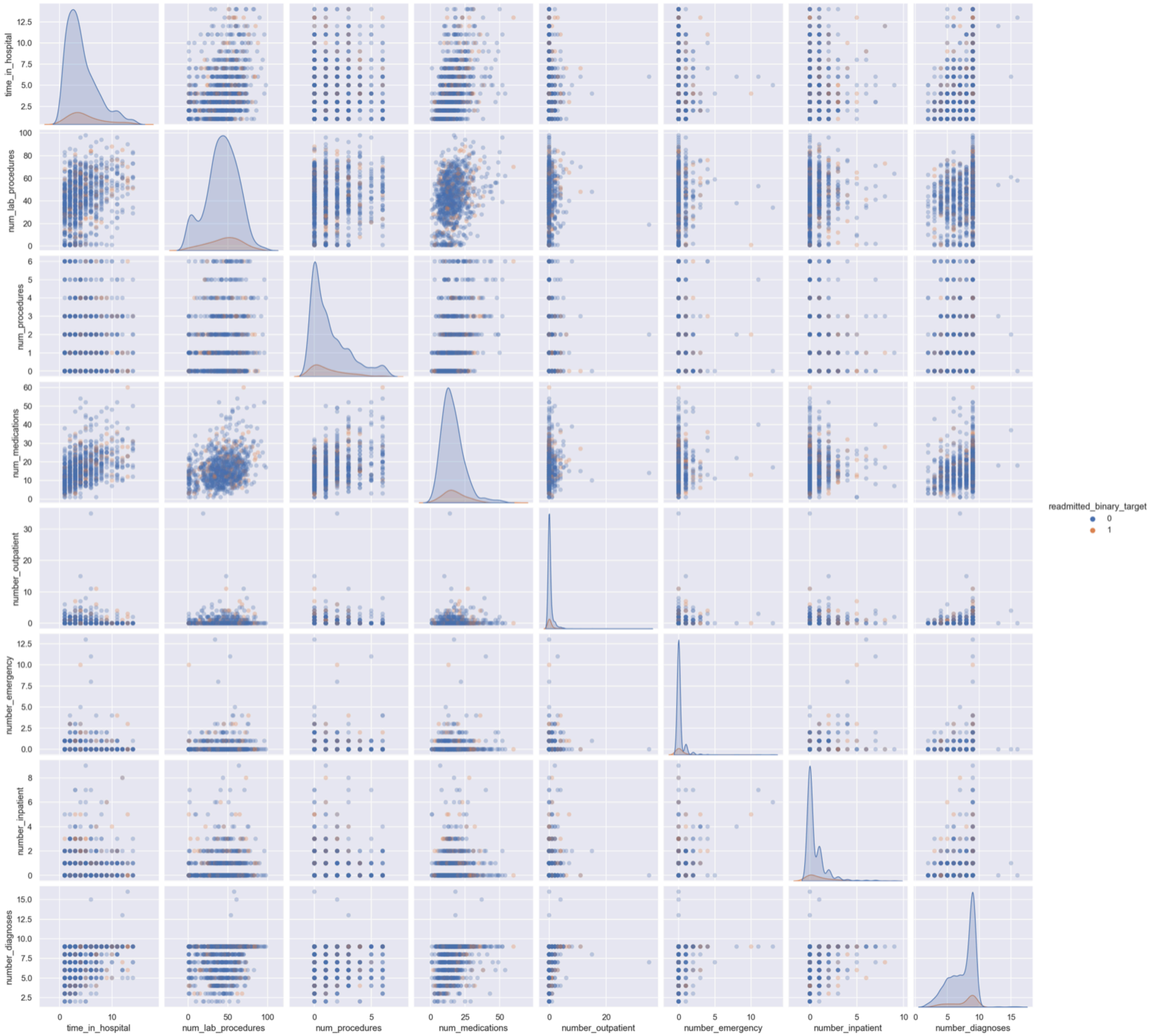
-**Data bleeding** was avoided by implementing a cross-validating scheme that prevented that data from the same patient appeared both in the training folds and the validation/testing.

-10 fold cross-validation.

-Imbalance was handled by a balanced class weighting multiplier. Oversampling and threshold adjustment to optimize F1 were tried But did not improve the F1 score.

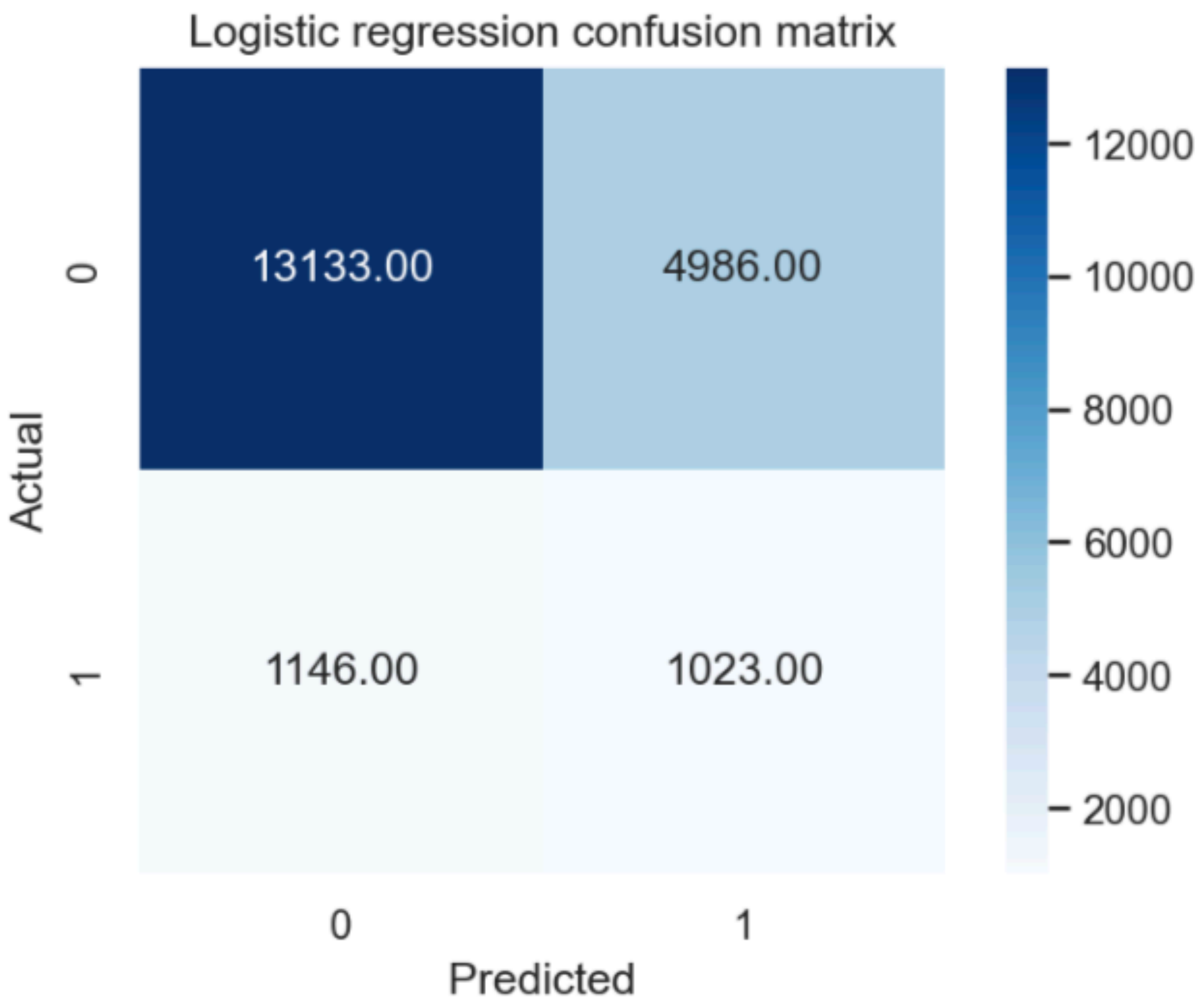
Appendix:

numerical features



Appendix:

Confusion matrix

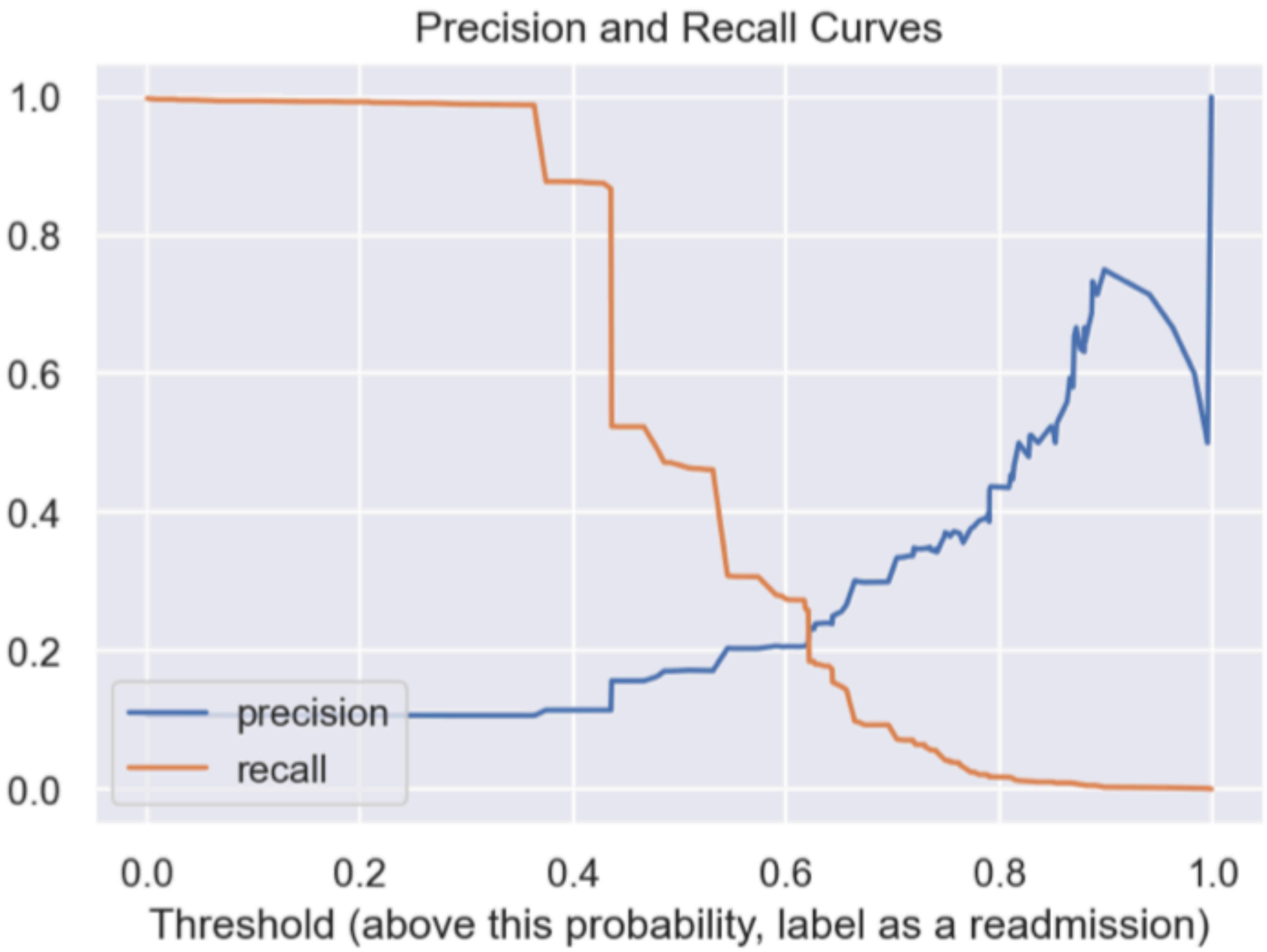


Logistic regression confusion matrix

Actual \ Predicted	0	1
0	13133.00	4986.00
1	1146.00	1023.00

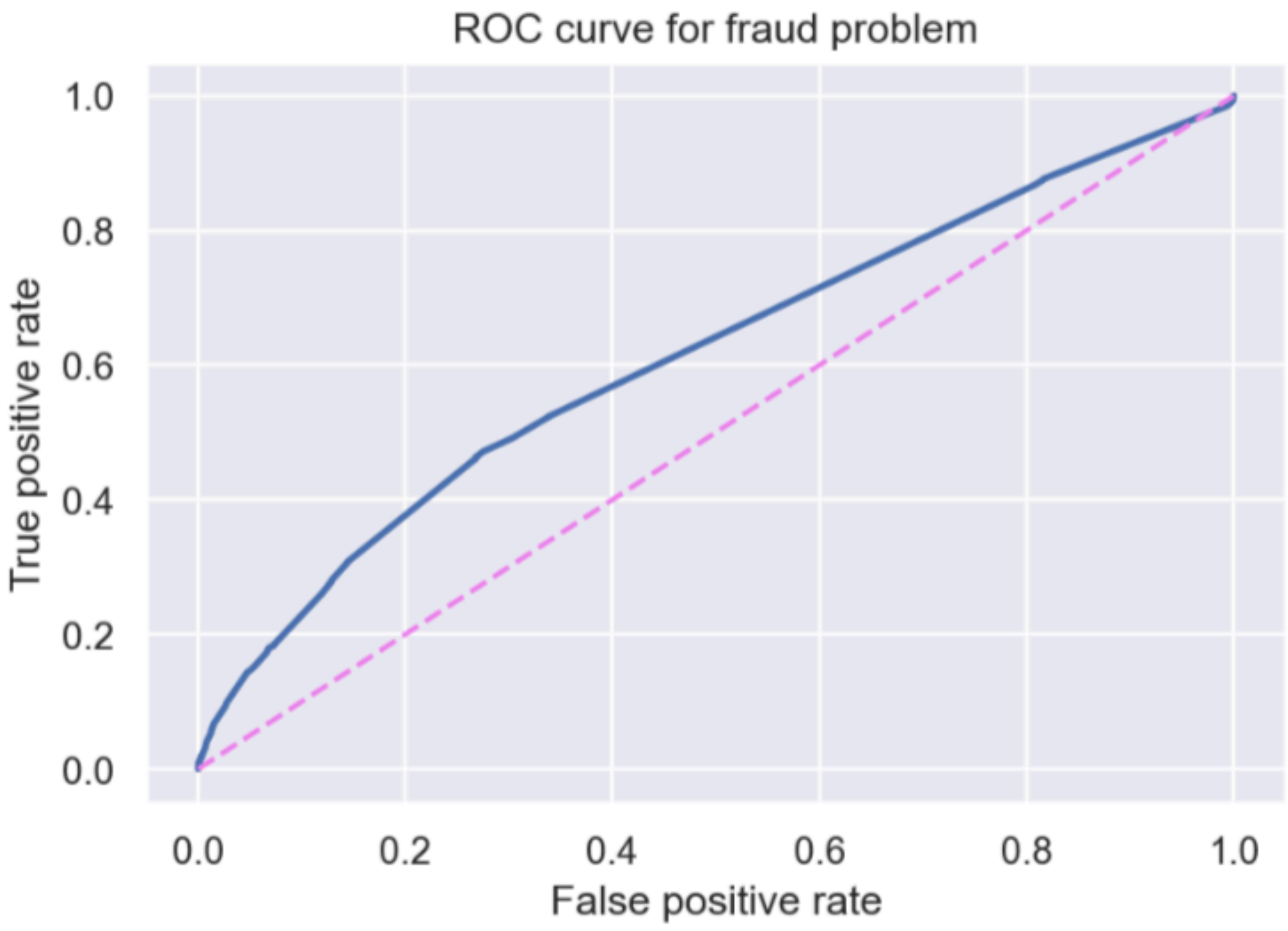
Appendix:

Precision recall-plot Random Forests

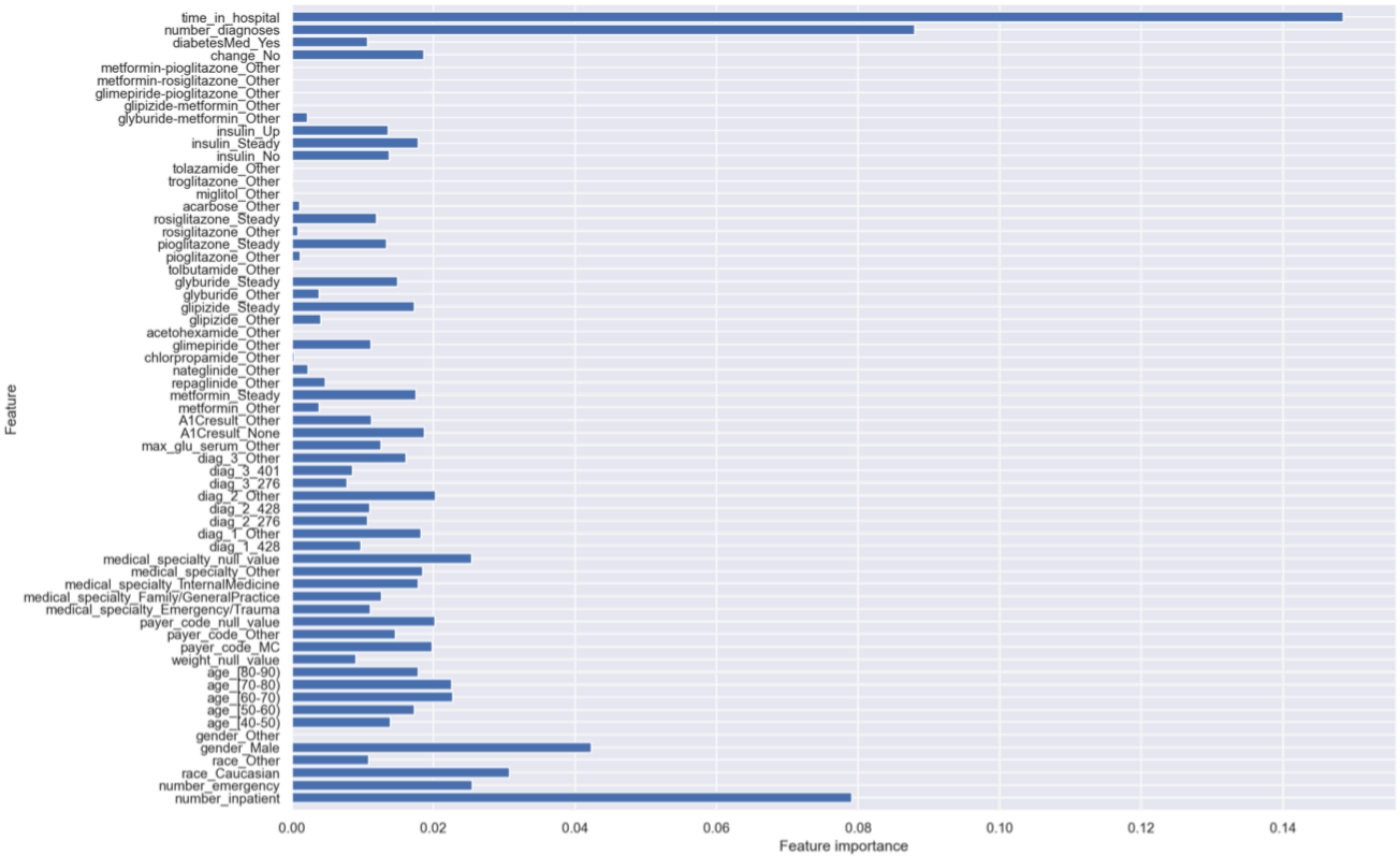


Appendix:

ROC Curve for Random Forests

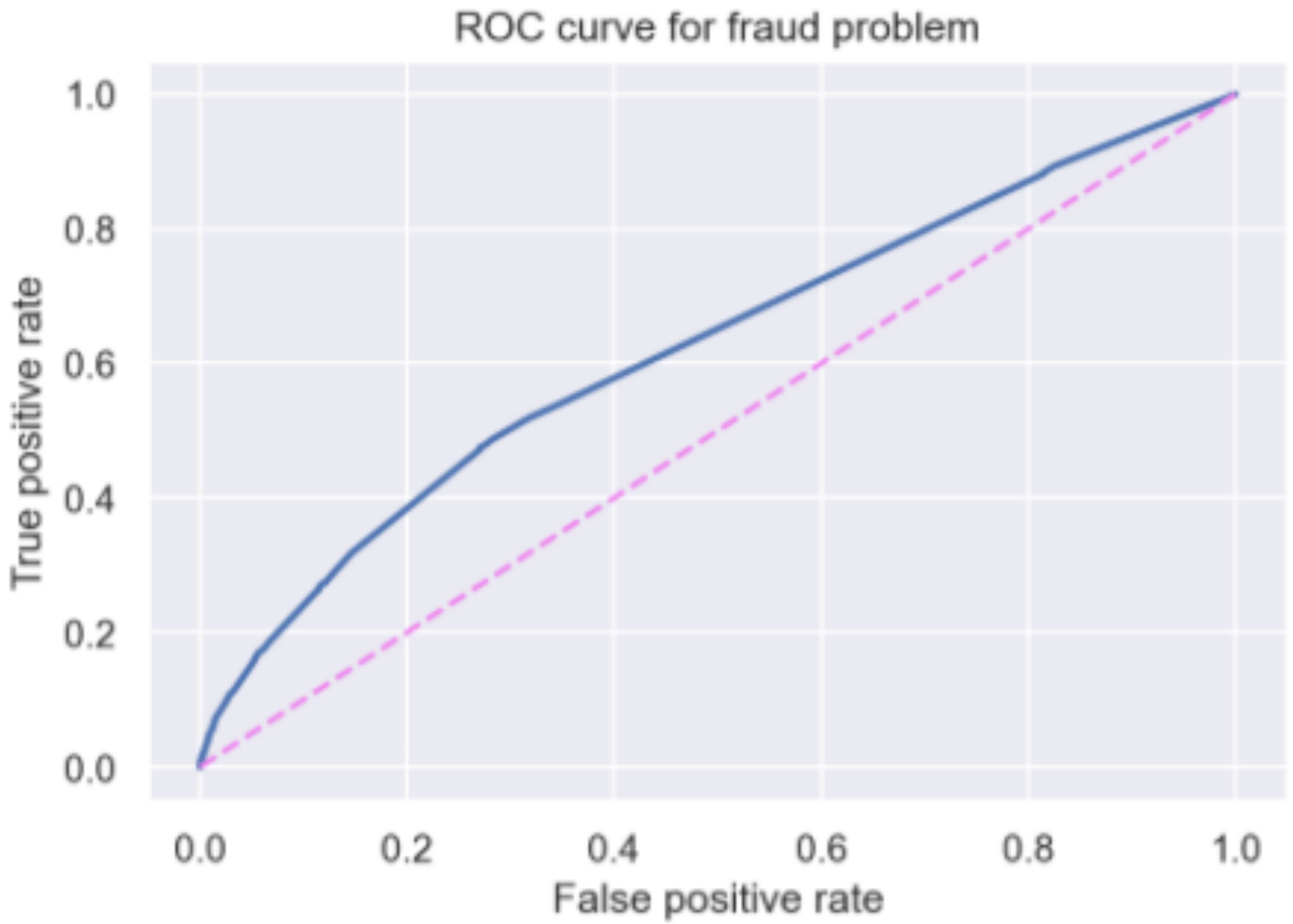


Random forests feature importance (10-fold cv)

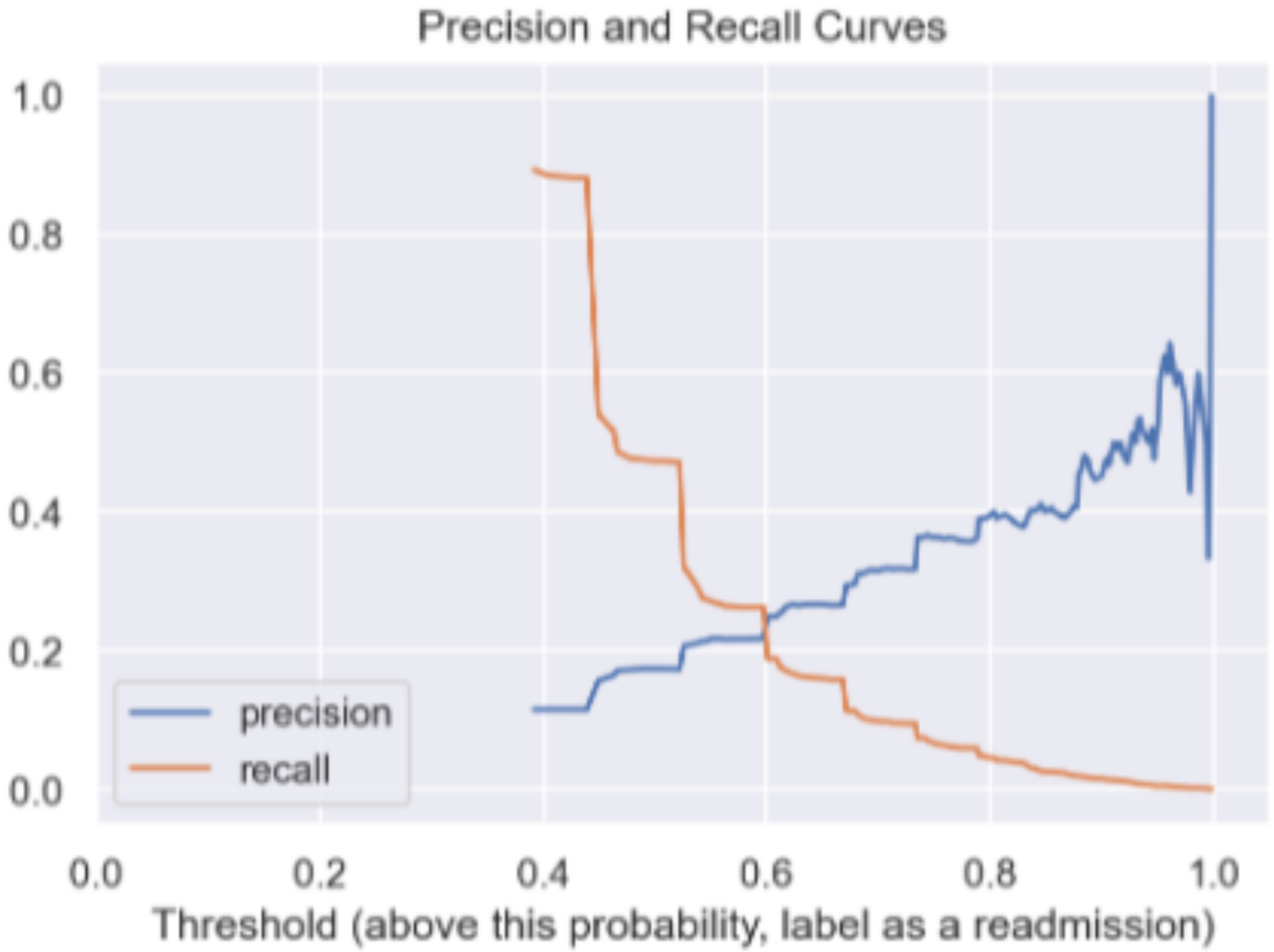


Appendix:

ROC Curve for Linear Regression



Precision recall-plotL Linear Regression


Precision and Recall Curves
precision
recall
Threshold (above this probability, label as a readmission)