

Predicting Diabetes Re-hospitalizations

By Maxim Belov

Diabetes in the US



- 34.2 million Americans (about 10%) have diabetes.
- \$237 billion were spent in 2017 on direct care.
- \$102 billion of that is related to direct hospital admissions.
- Patients with diabetes are more likely to be re-admitted following a hospitalization.
- What can we do to allow early identification and intervention for high-risk patients?

Agenda



Data Overview



Exploratory Data
Analysis

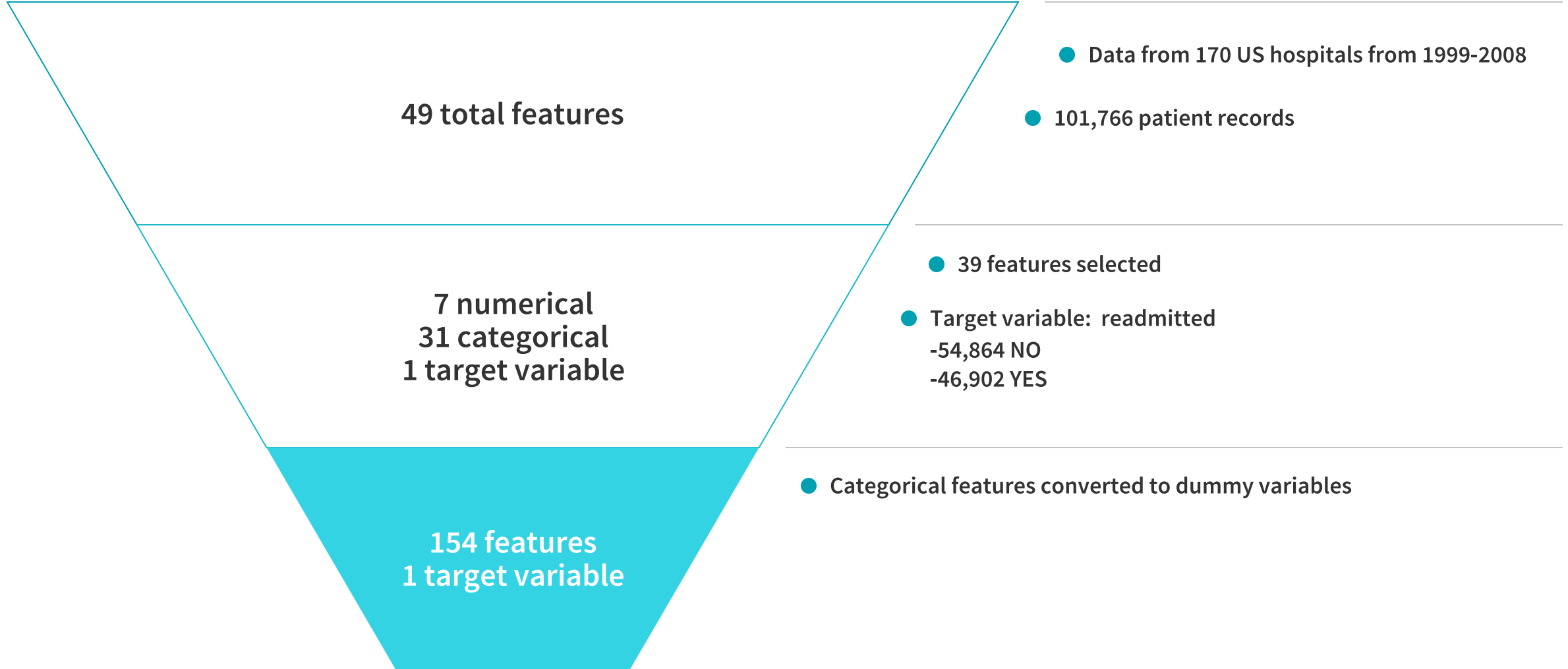


Supervised Machine
Learning Models

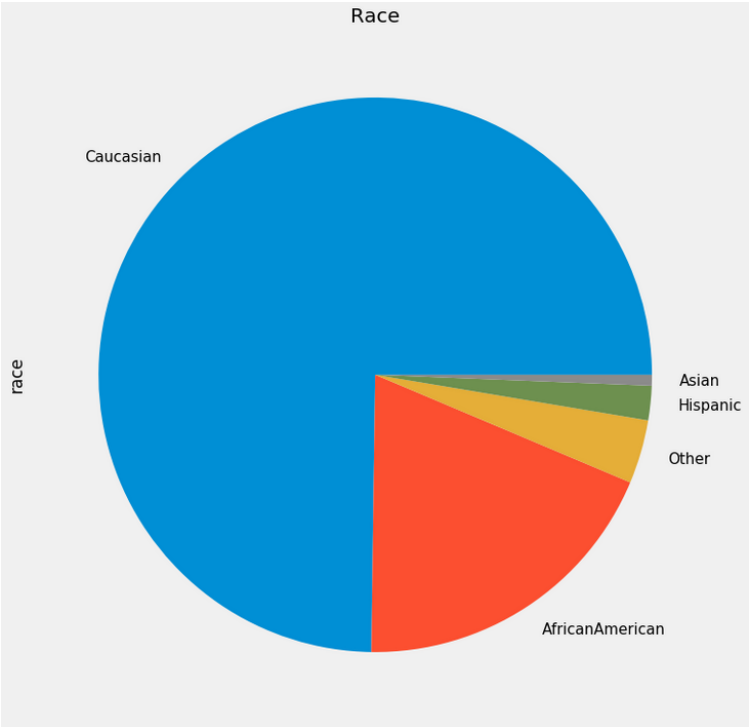
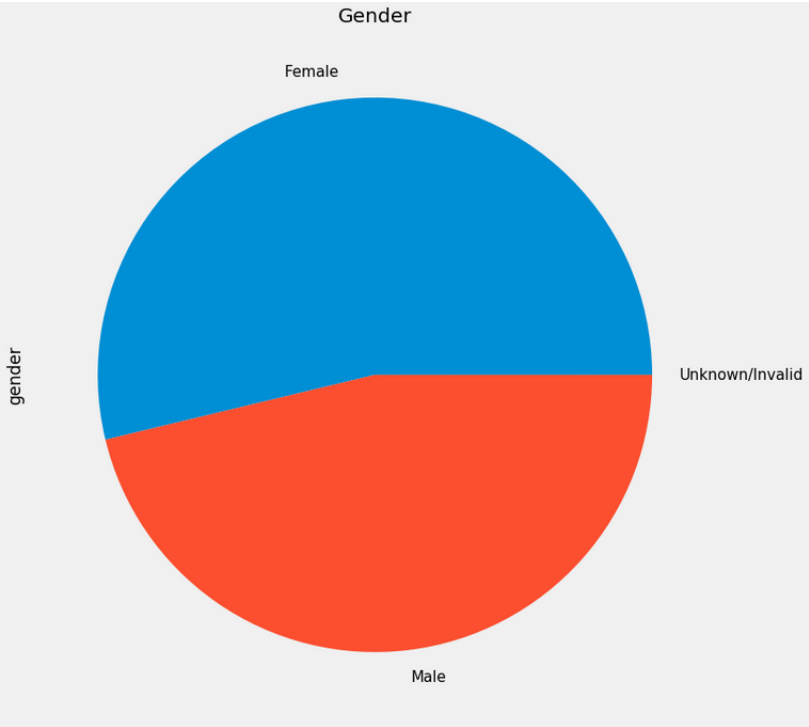
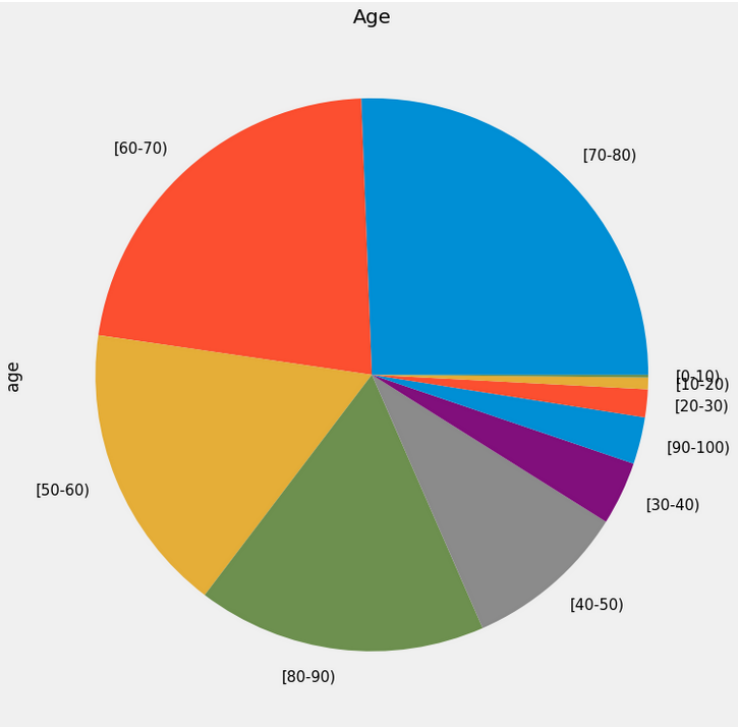


Results

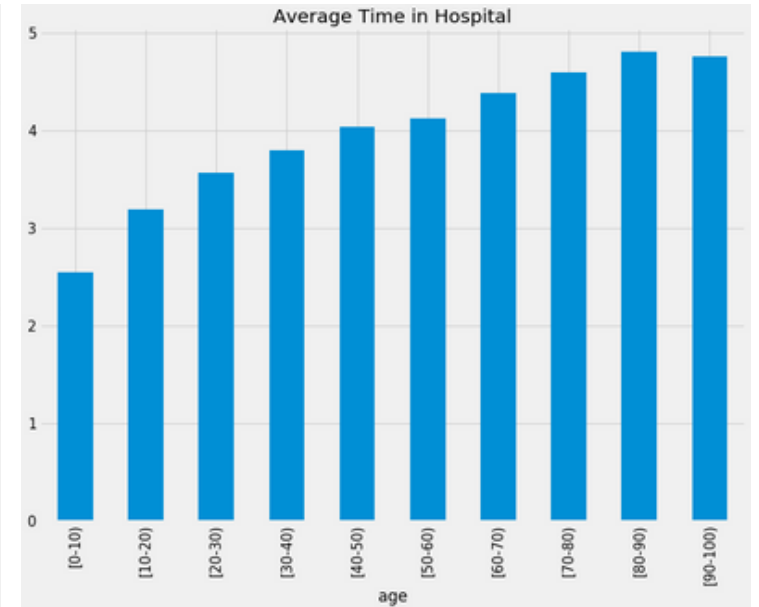
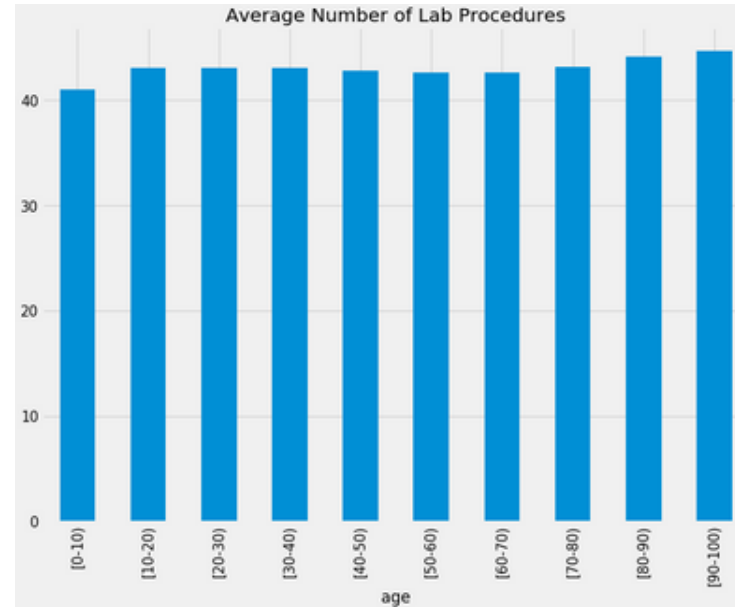
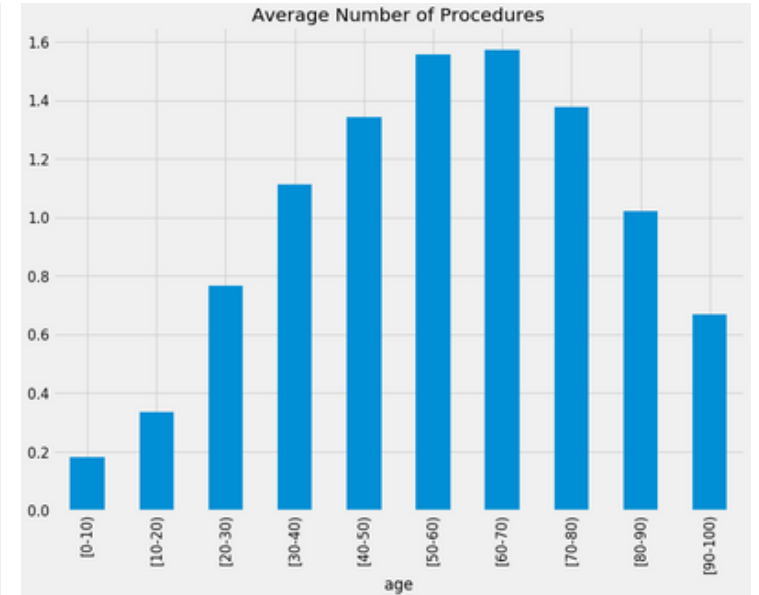
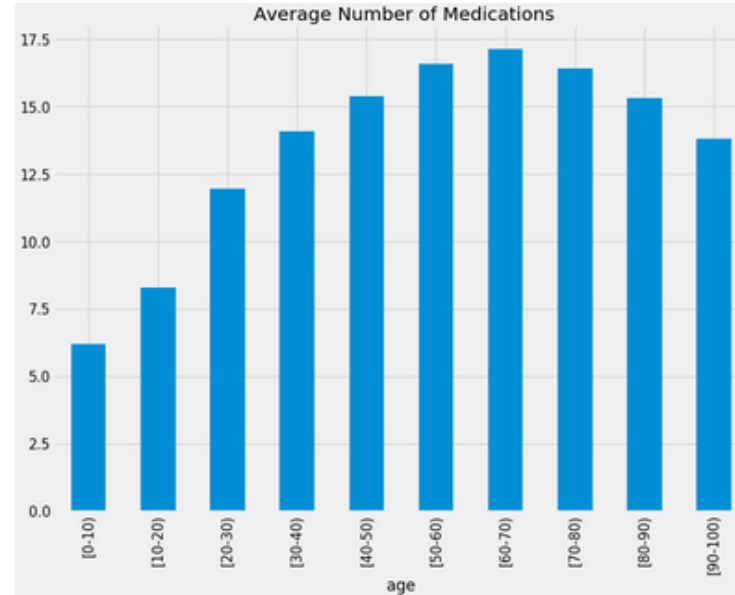
Data Overview



Exploratory Data Analysis

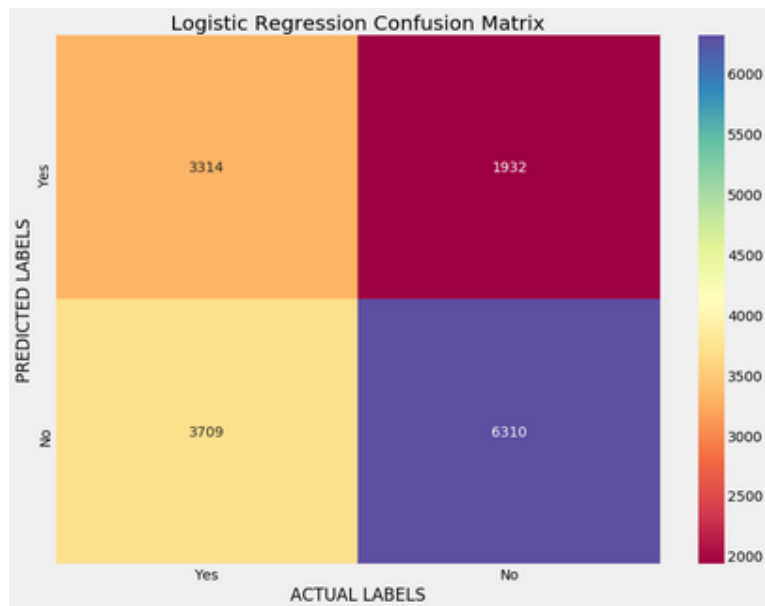


Exploratory Data Analysis



*Train-Test Split = 85/15

Base ML Models



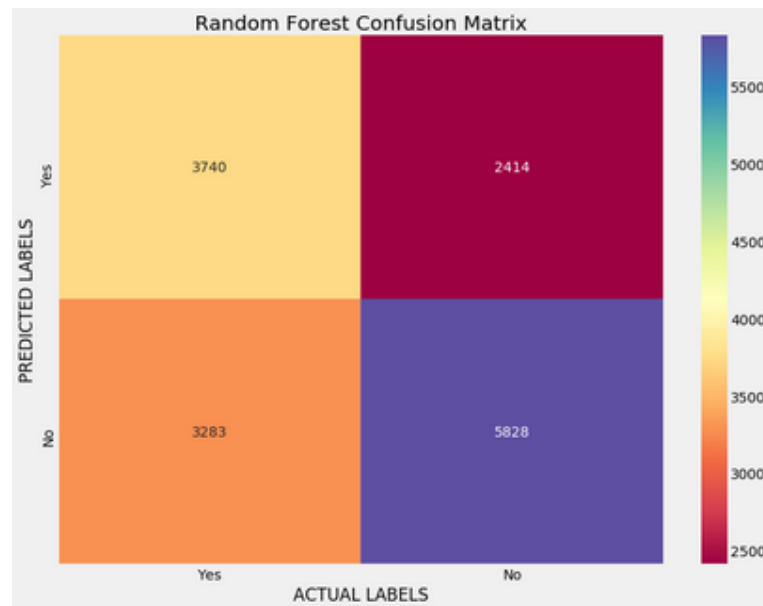
Logistic Regression

Accuracy: 0.63

Precision: 0.63

Recall: 0.62

F1: 0.62



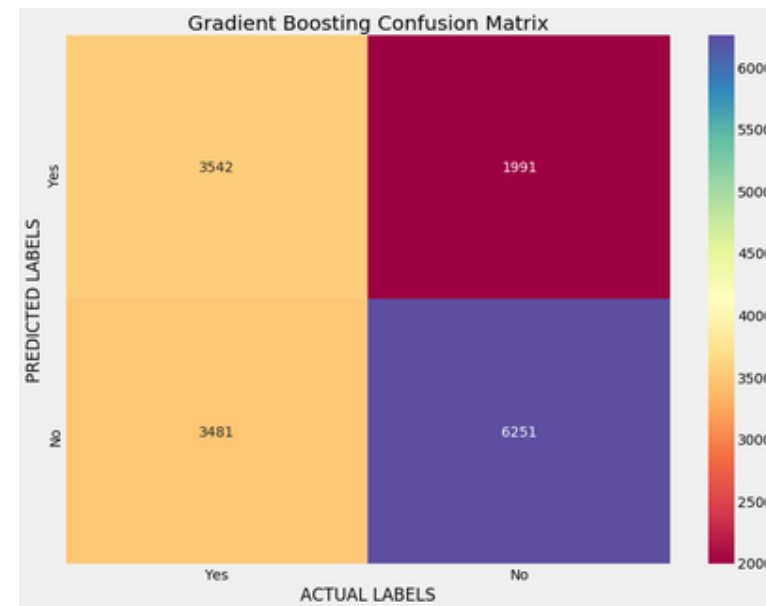
Random Forest Classifier

Accuracy: 0.64

Precision: 0.63

Recall: 0.62

F1: 0.62



Gradient Boosting Classifier

Accuracy: 0.65

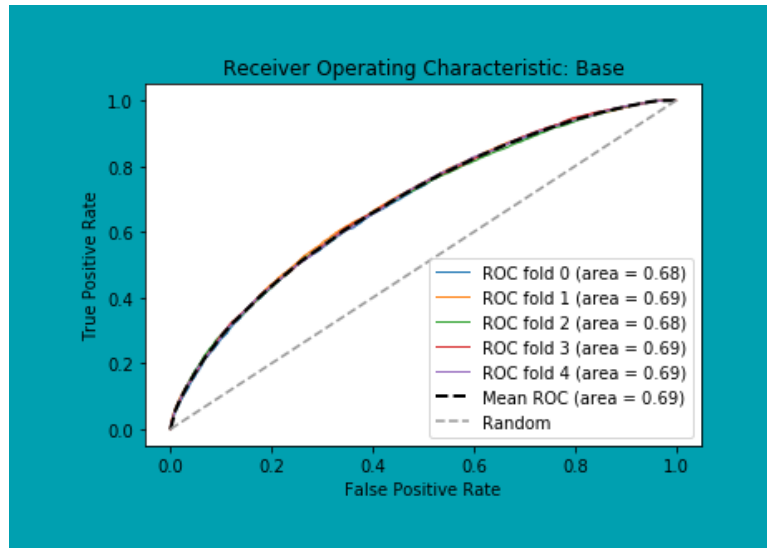
Precision: 0.64

Recall: 0.63

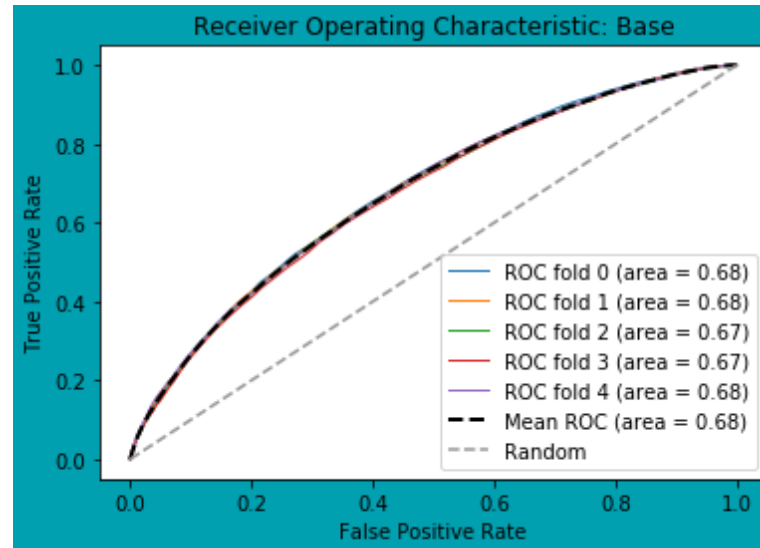
F1: 0.63

ROC Curves - Base

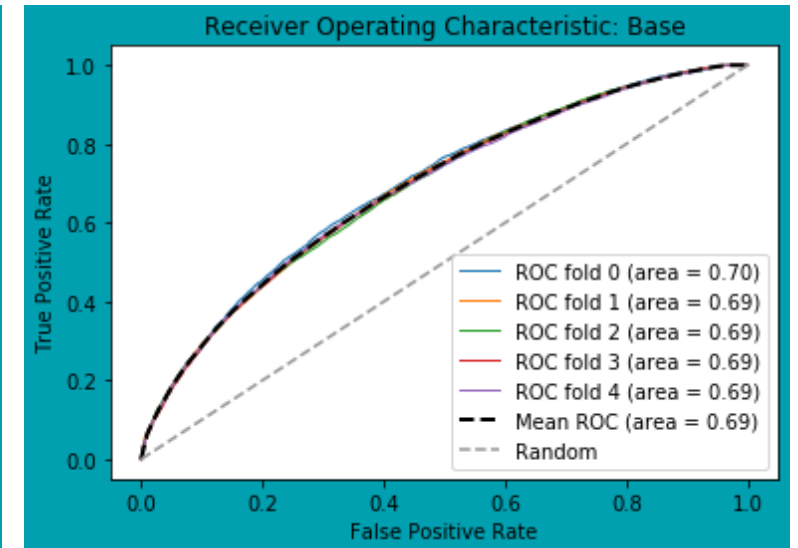
Logistic Regression



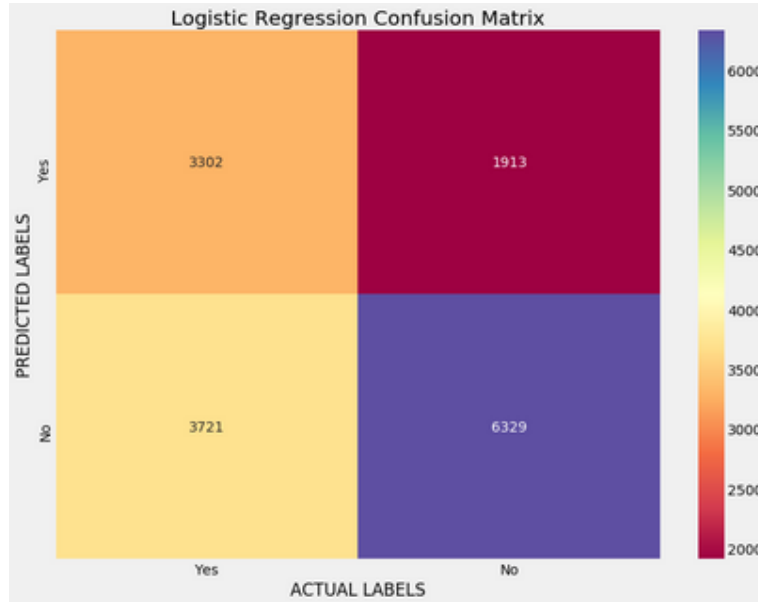
Random Forest



Gradient Boosting



Optimized ML Models



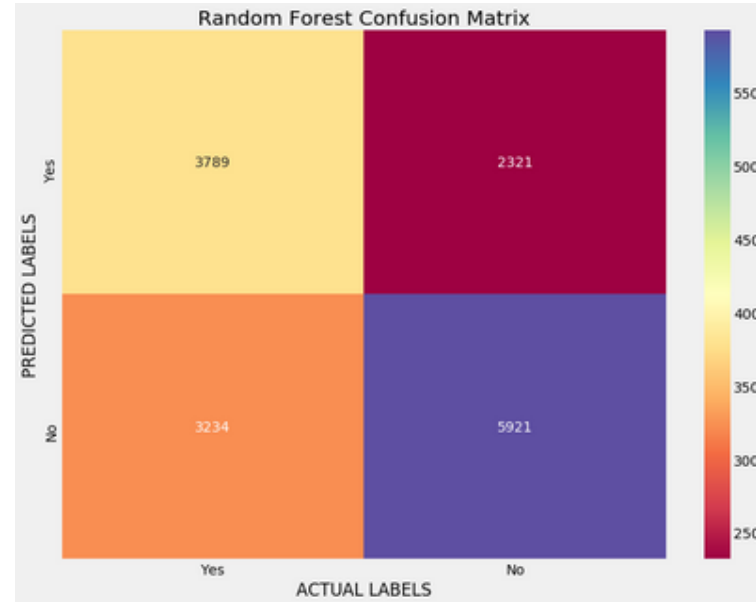
Logistic Regression

Accuracy: 0.63

Precision: 0.63

Recall: 0.62

F1: 0.62



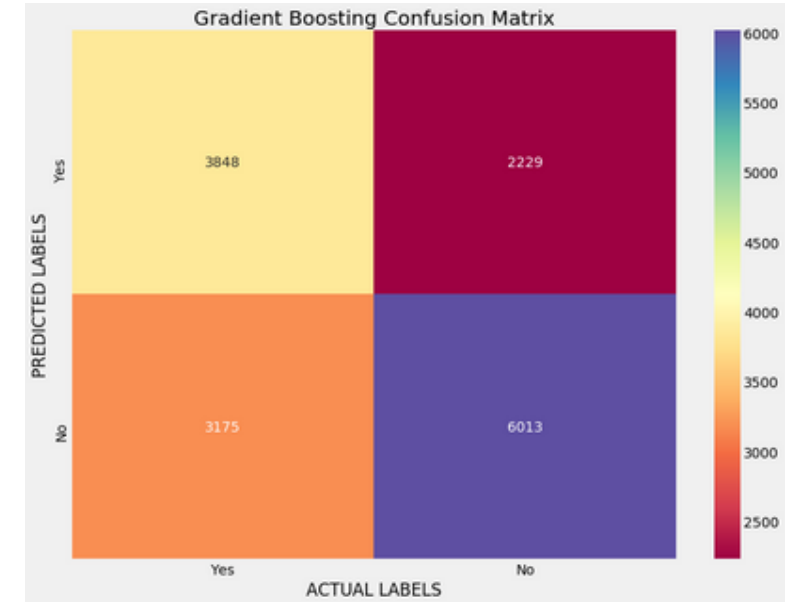
Random Forest

Accuracy: 0.64

Precision: 0.63

Recall: 0.63

F1: 0.63



Gradient Boosting

Accuracy: 0.65

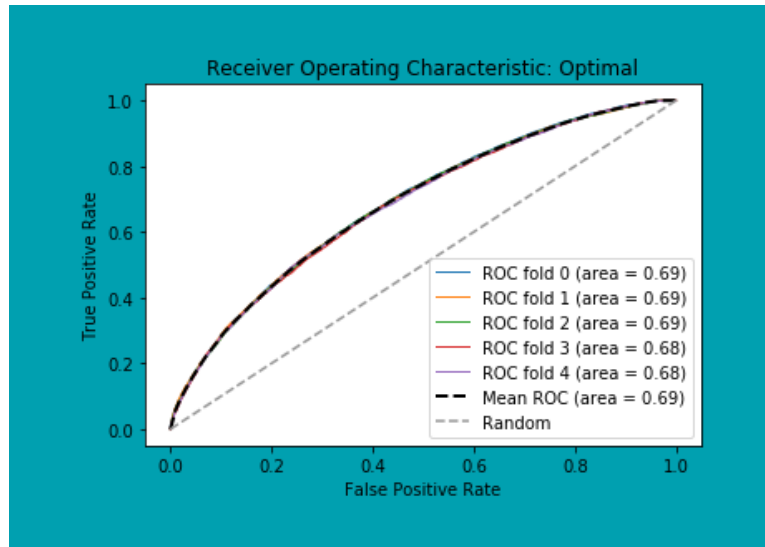
Precision: 0.64

Recall: 0.64

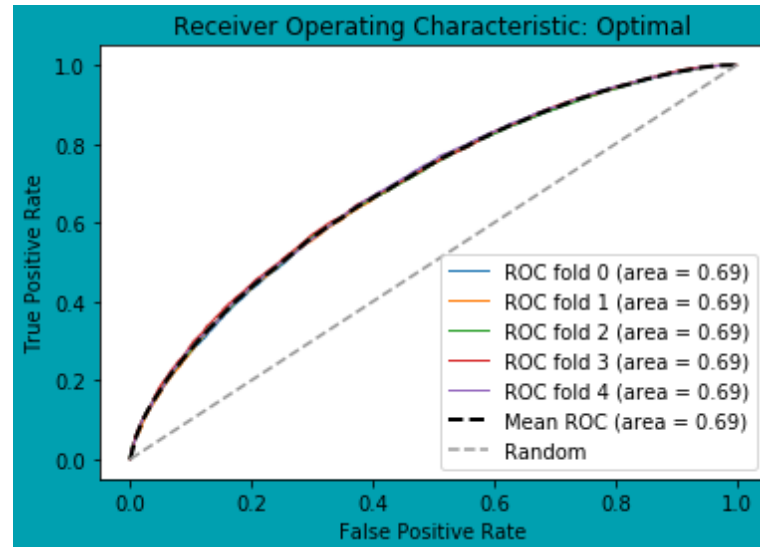
F1: 0.64

ROC Curves - Optimized

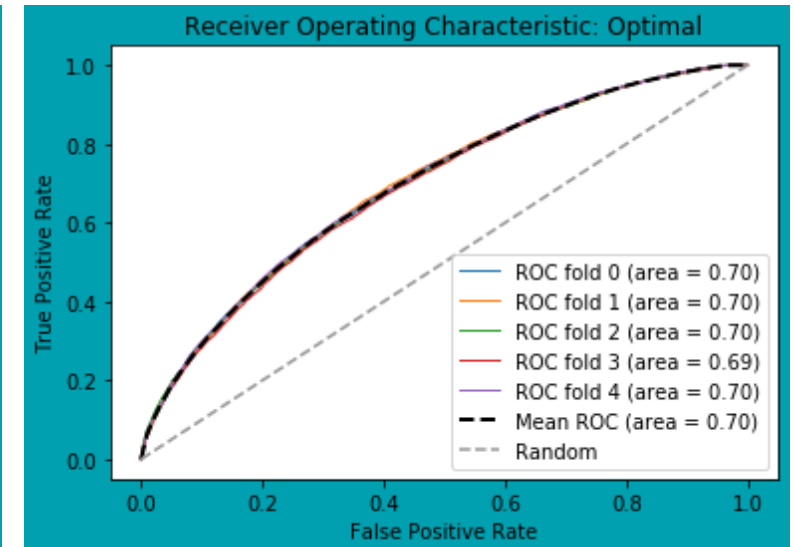
Logistic Regression



Random Forest



Gradient Boosting



Conclusion

- **Performance**

- All 3 models performed poorly out-of-the box.
- Minor improvements after optimizing hyper-parameters.
- Not ready for real world data.

- **Next steps**

- Review the data to see how it can be better organized.
- Remove certain features.
- Redefine target classes.
- Try different models.