Blood Donation Prediction: A Machine Learning Approach

Joseph Petersen

Introduction and Problem Statement

- Fictional Client: High Yield Platelet Obtainment (HYPO): A small non-profit that runs pop-up blood banks
- HYPO is seeking to predict return donators based off previous donation data.
- Not losing vital blood donation is deemed the most important factor given its constant medical necessity and shelf life.
- Dataset: https://archive.ics.uci.edu/ml/machine-learning-databases/blood-transfusion/

Dataset Description

- Response Variable:
  - donated_march_2007 - Binary (0 or 1) indicating whether the individual returned to donate in March, 2007
- Features:
  - months_since_last – Months since the last visit
  - months_since_first – Months since the first visit
  - total_times_donated – Total number of visits
  - total_cc_blood_donated – Total amount of blood donated in cubic centimeters
- Total entries: 748
- Distribution of the response variable: ~24%

Methodology

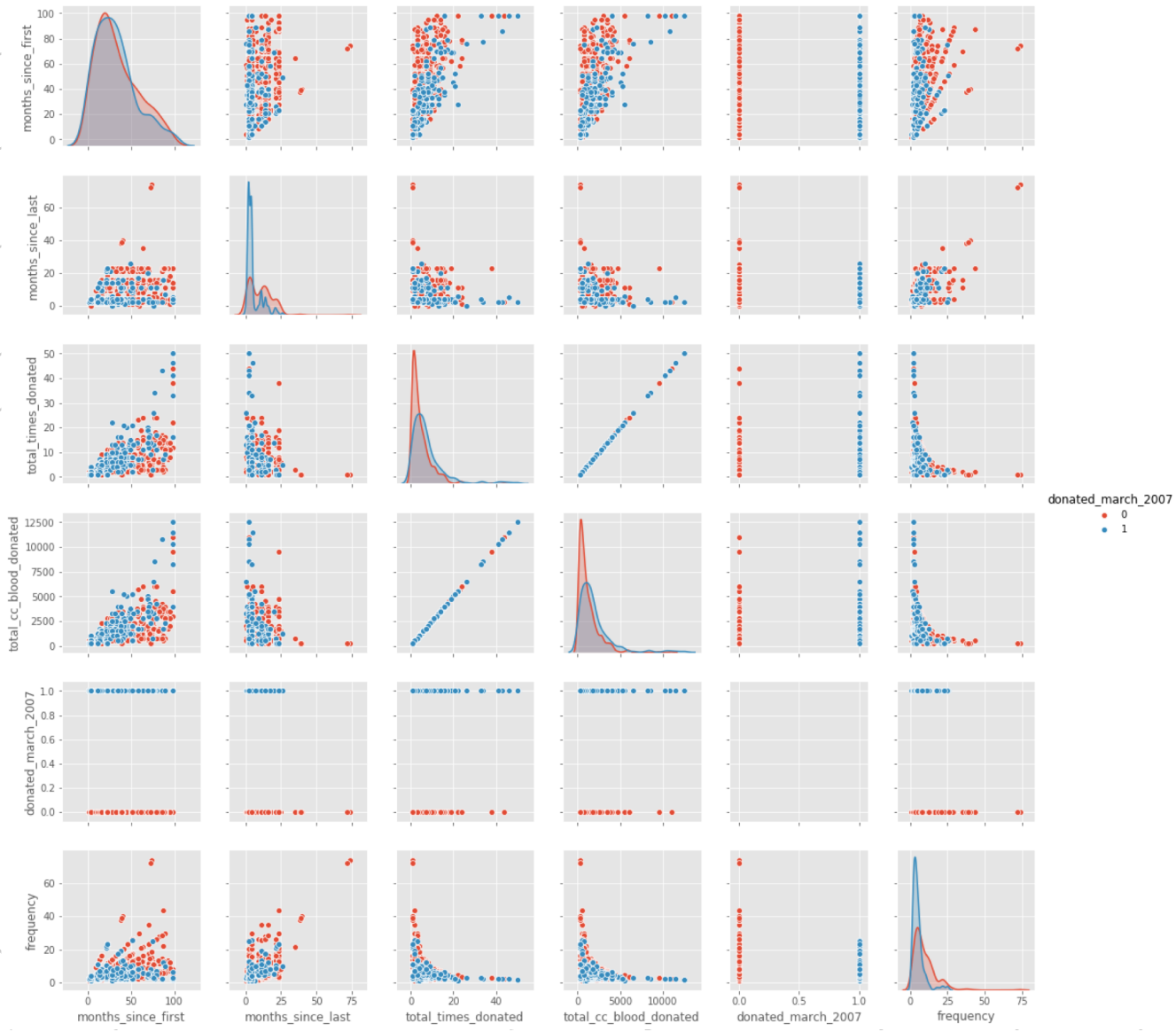
- 1. Explore the dataset and determine issues with the data or outliers
- 2. Perform exploratory data analysis to determine strong predictors
- 3. Scale and upsample the data before fitting and testing appropriate machine learning models (This is a classification problem so specifically classifiers)
- Note: Recall will be an important scoring metric as the client wishes to minimize loss of potential donators (False Negatives)

Data Wrangling and Initial Impressions

- The data was delivered clean.
- The features were not comparable in scale.
- Outliers were detected in 'months_since_last', 'total_times_donated', and 'total_cc_blood_donated'.
- 'total_times_donated' and 'total_cc_blood_donated' appear to be a product of the other.
- 'frequency' was calculated as 'months_since_first' / 'total_times_donated'.

Hypotheses and Results

- All tested with $\alpha < 0.05$ using t-tests.
- Is frequency of visits a significant factor on returning?
  - H_0: There is no difference in mean frequency.
  - H_A: Returners (1) visit more frequently (lower mean frequency).
- Is time since last visit a significant factor on returning?
  - H_0: There is no difference in mean time since last visit.
  - H_A: There is a difference.
- Is total number of visits a significant factor on returning?
  - H_0: There is no difference in mean total number of visits.
  - H_A: There is a difference.


months_since_first
months_since_last
total_times_donated
total_cc_blood_donated
donated_march_2007
frequency
donated_march_2007
0
1

Determinations Based on the Results

- 'total_cc_blood_donated' is indeed a product of 'total_times_visited' given the direct linear correlation.
- 'frequency' and 'months_since_last' are both strong predictors.
- 'frequency' as a product of 'total_times_donated' and 'months_since first' could be a better predictor with those features removed (curse of dimensionality).
- Some correlation between features besides

Methodology for Machine Learning

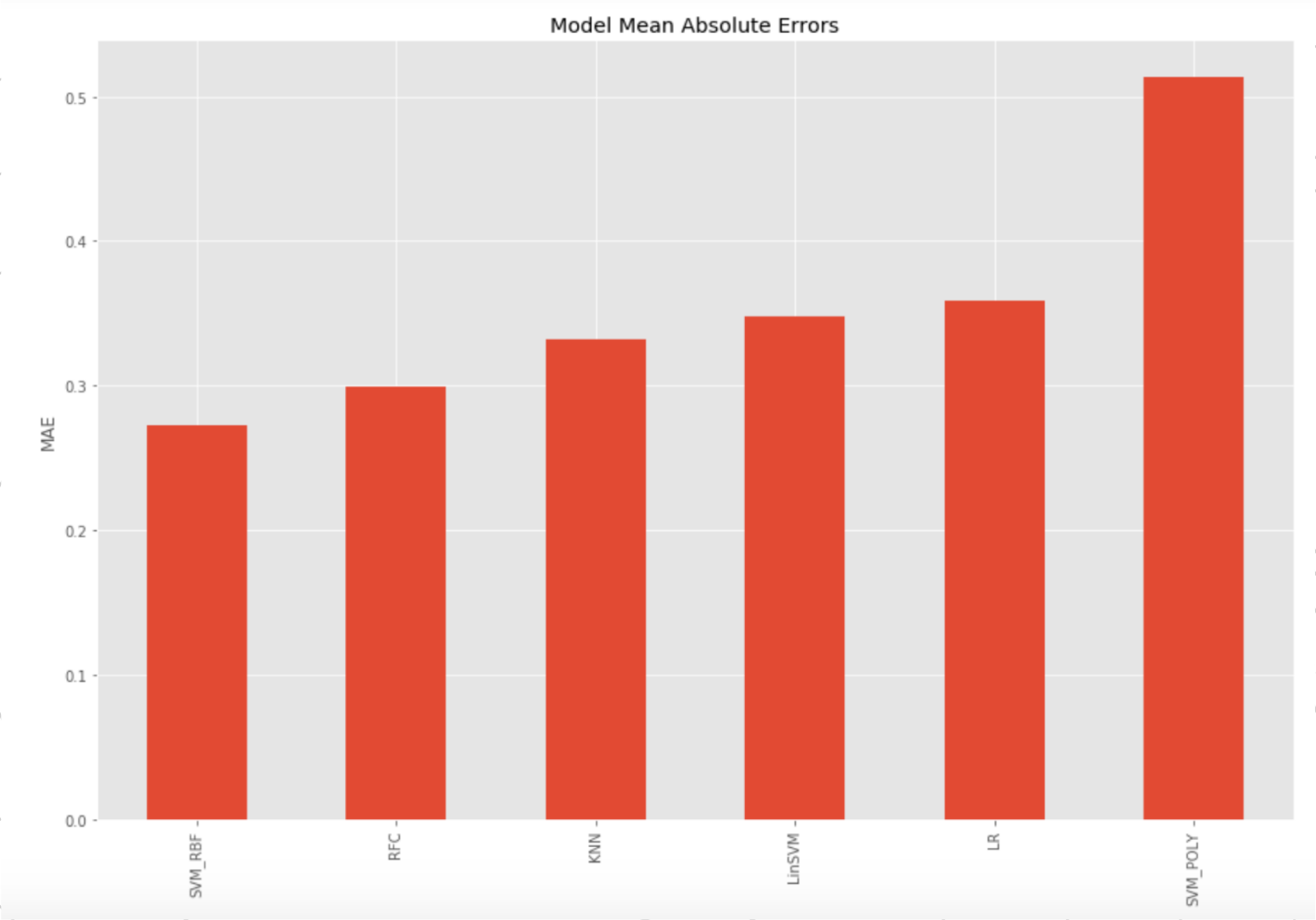
- Models to be tested:
  - Logistic Regression: Don't need to worry as much about correlations between variables. Can easily adjust thresholds based on probability.
  - Support Vector Machine: Different kernels allow flexibility on data not linearly separated.
  - Random Forest: Don't need to be concerned as much with outliers and non-linearity as it's non-parametric.
  - K-Nearest Neighbors: Can deal with noisy data.

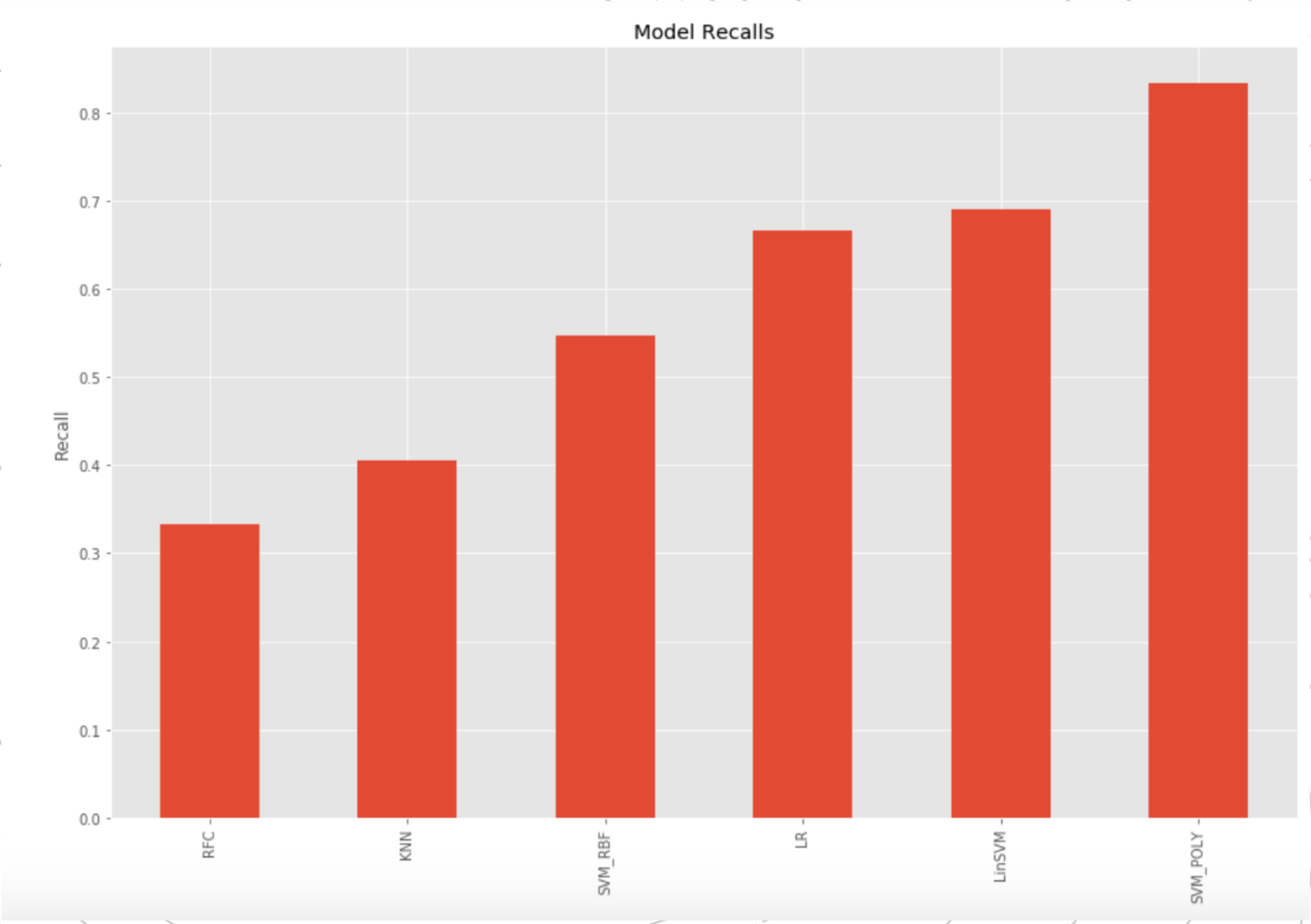
Methodology for Machine Learning

- The hypothesis is that the minimized dataset ('frequency', 'months_since_last') will be the most effective given the reduced dimensionality and strength of the predictors.
- However, both that and the other ('frequency', 'months_since_last', 'months_since_first', 'total_times_donated') will be tested to determine that hypothesis.
- The data will be scaled using RobustScaler (good with outliers as it scales to the interquartile range) and upsampled using SMOTE (response variable: ~24% distribution)

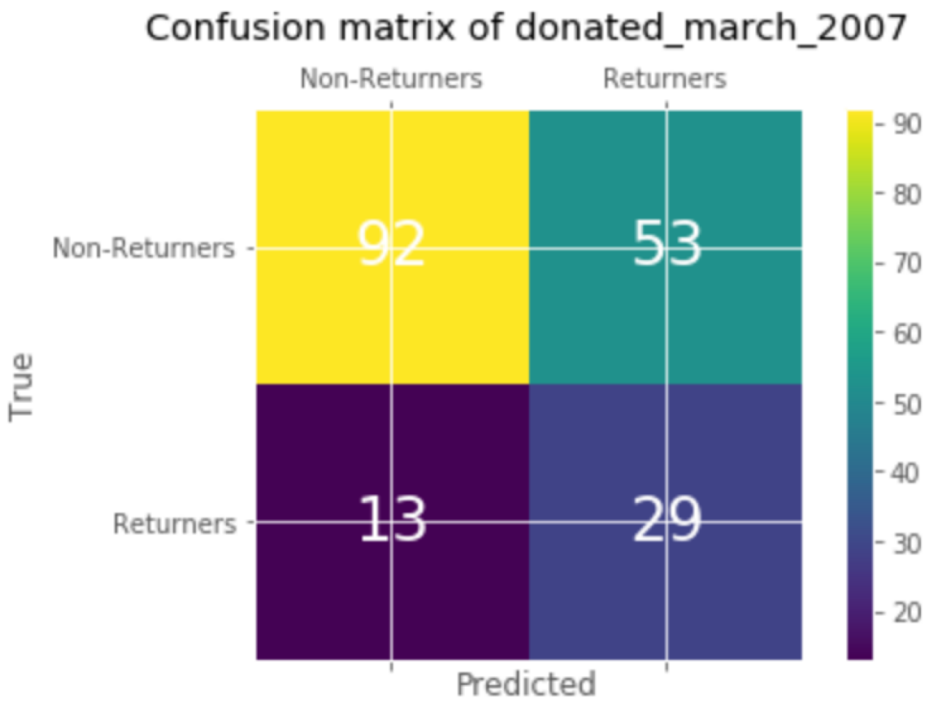
Score Table for Larger Dataset

	accuracy	precision	recall	f1	mae
LR	0.641711	0.345679	0.666667	0.455285	0.358289
LinSVM	0.652406	0.358025	0.690476	0.471545	0.347594
RFC	0.700535	0.333333	0.333333	0.333333	0.299465
SVM_RBF	0.727273	0.418182	0.547619	0.474227	0.272727
SVM_POLY	0.486631	0.282258	0.833333	0.421687	0.513369
KNN	0.668449	0.314815	0.404762	0.354167	0.331551


Model Mean Absolute Errors
MAE
0.0
0.1
0.2
0.3
0.4
0.5
SVM_RBF
RFC
KNN
LinSVM
LR
SVM_POLY


Model Recalls
Recall
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
RFC
KNN
SVM_RBF
LR
LinSVM
SVM_POLY

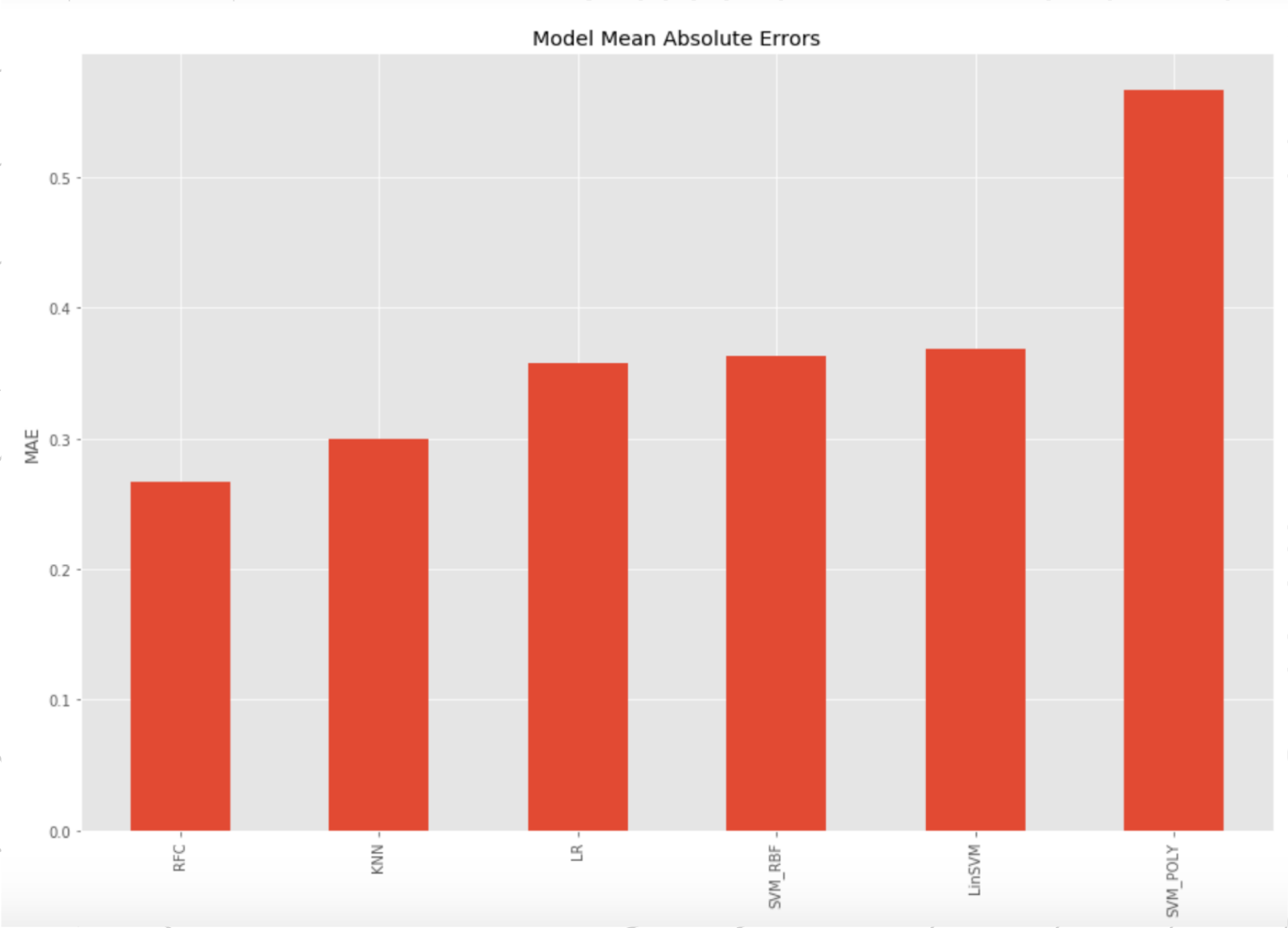
Tuning the Linear Support Vector Machine Model

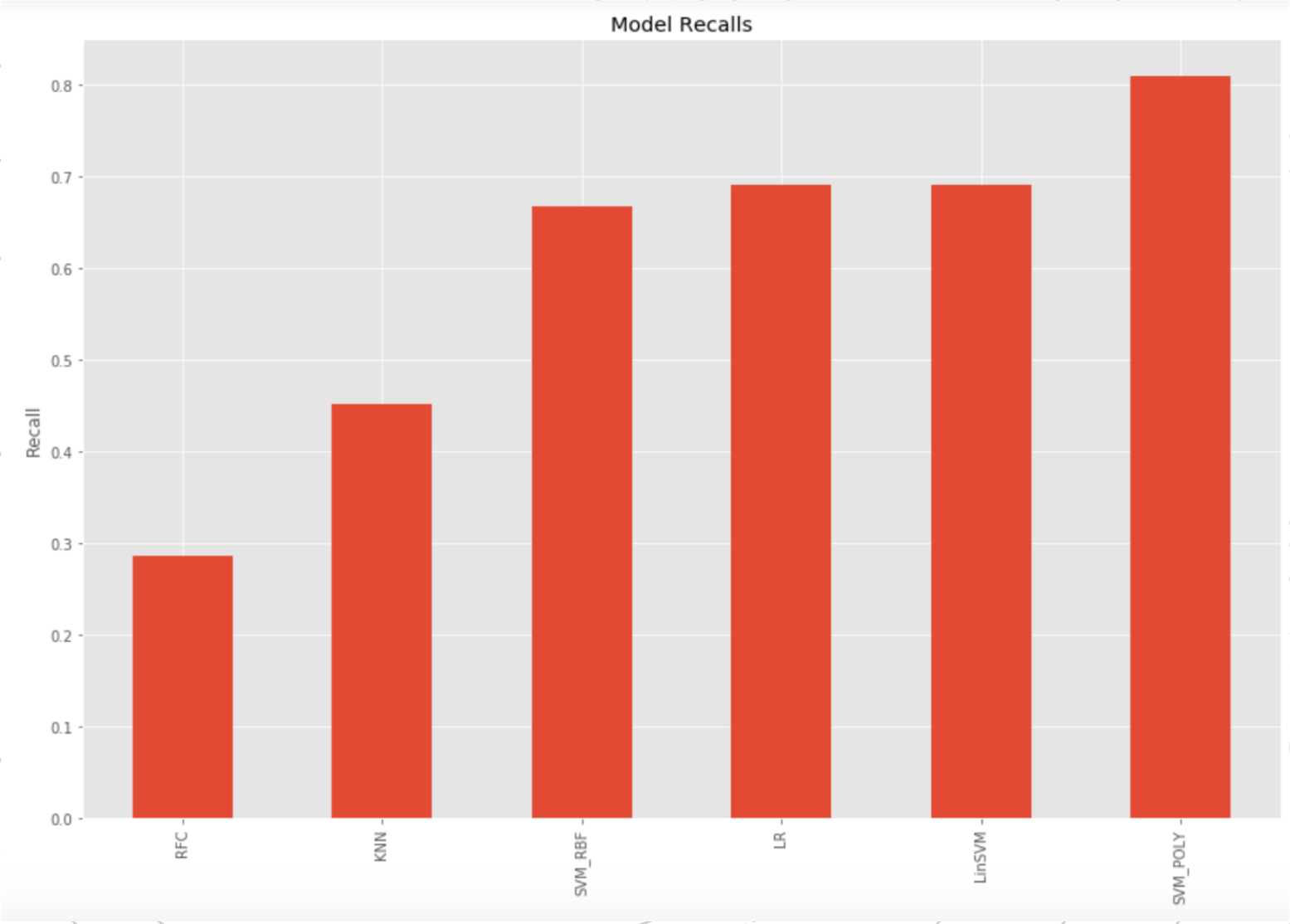


	precision	recall	f1-score	support
0	0.88	0.63	0.74	145
1	0.35	0.69	0.47	42
micro avg	0.65	0.65	0.65	187
macro avg	0.61	0.66	0.60	187
weighted avg	0.76	0.65	0.68	187

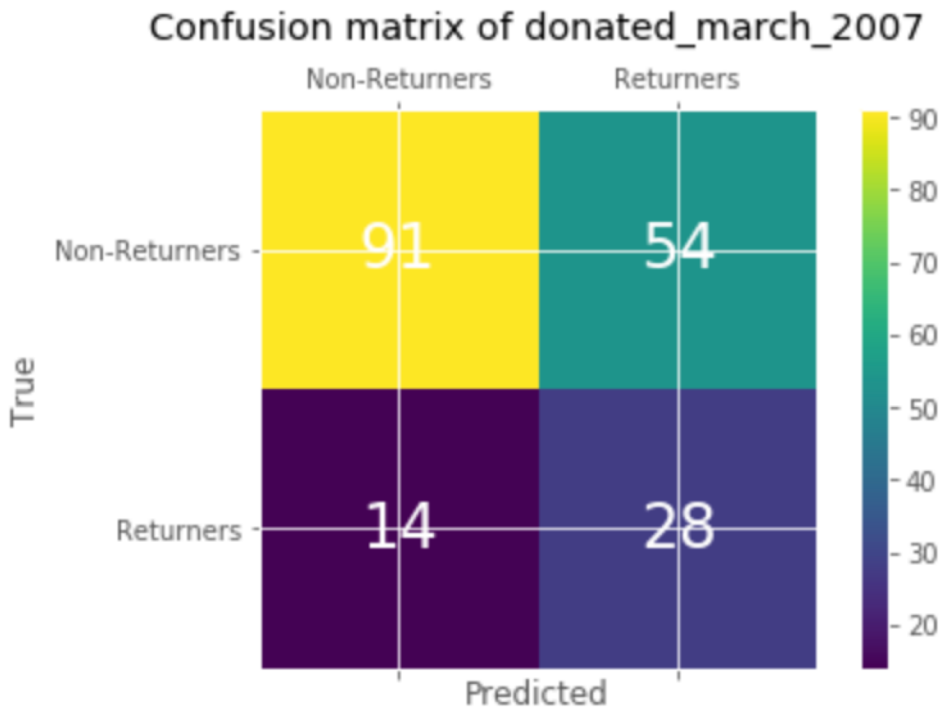
Score Table for Smaller Dataset

	accuracy	precision	recall	f1	mae
LR	0.641711	0.349398	0.690476	0.464	0.358289
LinSVM	0.631016	0.341176	0.690476	0.456693	0.368984
RFC	0.73262	0.375	0.285714	0.324324	0.26738
SVM_RBF	0.636364	0.341463	0.666667	0.451613	0.363636
SVM_POLY	0.433155	0.257576	0.809524	0.390805	0.566845
KNN	0.700535	0.365385	0.452381	0.404255	0.299465


Model Mean Absolute Errors
MAE
0.0
0.1
0.2
0.3
0.4
0.5
RFC
KNN
LR
SVM_RBF
LinSVM
SVM_POLY

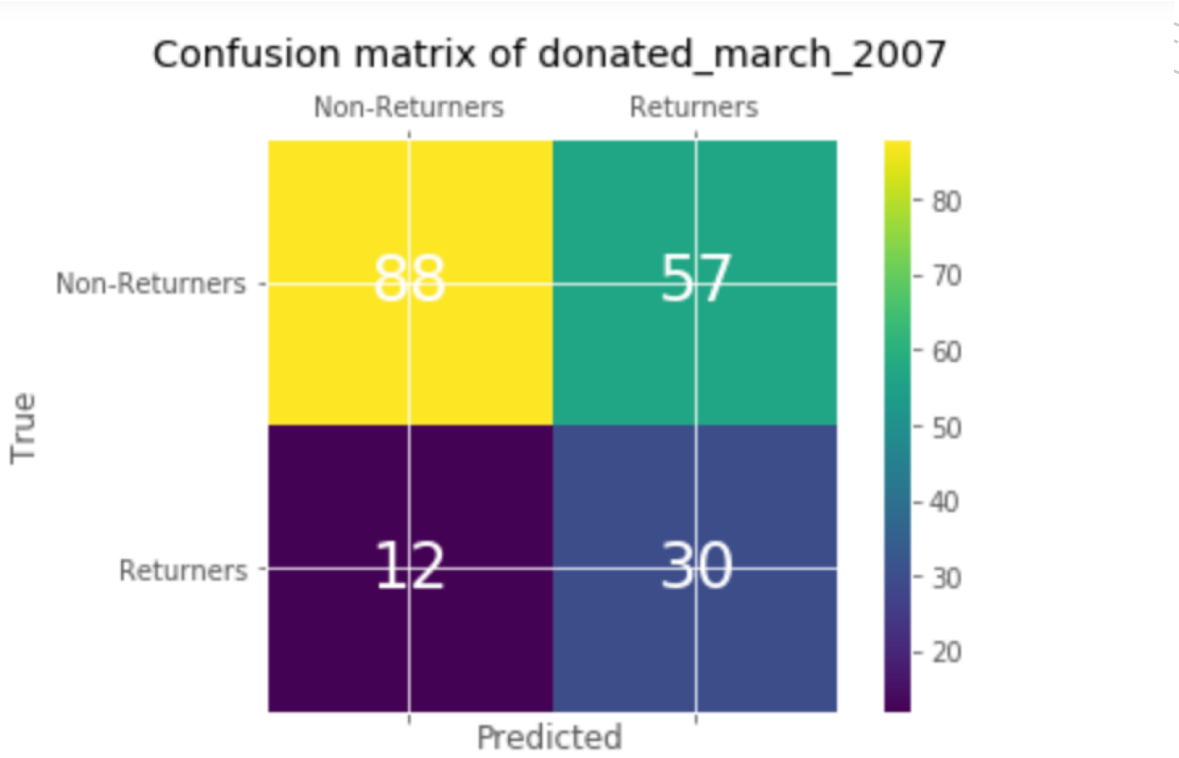

Model Recalls
Recall
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
RFC
KNN
SVM_RBF
LR
LinSVM
SVM_POLY

Tuning the Logistic Regression Model on Reduced Dataset



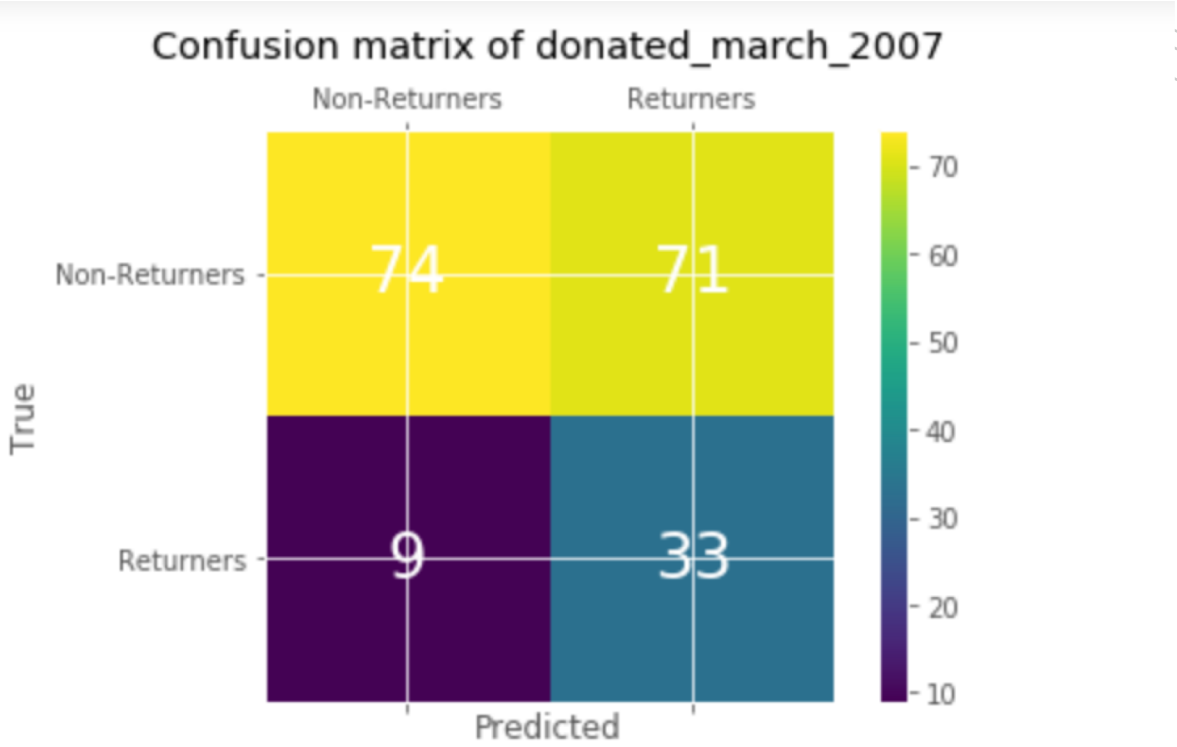
	precision	recall	f1-score	support
0	0.87	0.63	0.73	145
1	0.34	0.67	0.45	42
micro avg	0.64	0.64	0.64	187
macro avg	0.60	0.65	0.59	187
weighted avg	0.75	0.64	0.67	187

Adjusting the Probabilistic Threshold (0.47)



	precision	recall	f1-score	support
0	0.88	0.61	0.72	145
1	0.34	0.71	0.47	42
micro avg	0.63	0.63	0.63	187
macro avg	0.61	0.66	0.59	187
weighted avg	0.76	0.63	0.66	187

Adjusting the Probabilistic Threshold (0.40)



	precision	recall	f1-score	support
0	0.89	0.51	0.65	145
1	0.32	0.79	0.45	42
micro avg	0.57	0.57	0.57	187
macro avg	0.60	0.65	0.55	187
weighted avg	0.76	0.57	0.60	187

Thoughts for Client

- The threshold for prediction is very sensitive. Further data collection could strengthen the model.
- The model as it stands does minimize false negatives so missing out on possible donators is also minimized.
- Marketing material and information on future blood drives could prove effective on false positives.