PREDICTING PATIENT APPOINTMENT NO-SHOWS

CATHERINE MAGSINO

COST OF NO-SHOWS

Up To
30%
No-show rates nationwide

$200
Average cost per unused appt slot

$150B
per year for the US healthcare system

More importantly, these unused appointments could be going to patients that need them.

OBJECTIVE

To create a model that predicts whether or not a patient will show up to an appointment, and recommend a course of action for the identified individuals.

METHODOLOGY

DATA

SOURCE: Kaggle

WHAT: 111K appointments, 62K Distinct Patients

WHERE: Across 81 neighborhoods

WHEN: April – June 2016

TOOLS

python

matplotlib

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

Seaborn

scikit learn

FEATURES

Classification: No-Show vs Show

Features:

- Age
- Gender
- Schedule Day of the Week
- Appointment Day of the Week
- Days Between Scheduling and Appt
- Number of Previous Appointments
- Number of Missed Appointments
- SMS Received
- Scholarship (Welfare)
- Handicap
- Diabetes
- Alcoholism
- Hypertension

Normalized using StandardScaler

20/80 Split Balanced using Random Oversampling

MODELS

Model	Balanced Accuracy
Logistic Regression	0.937
Random Forest	0.920
Decision Tree	0.886
XGBoost	0.945
K Nearest Neighbors	0.864
Support Vector Machine	0.941
SVM GridSearchCV	0.856

XGBOOST RESULT

Balanced Accuracy: 0.95

Recall: 0.99
Precision: 0.72
Specificity: 0.90

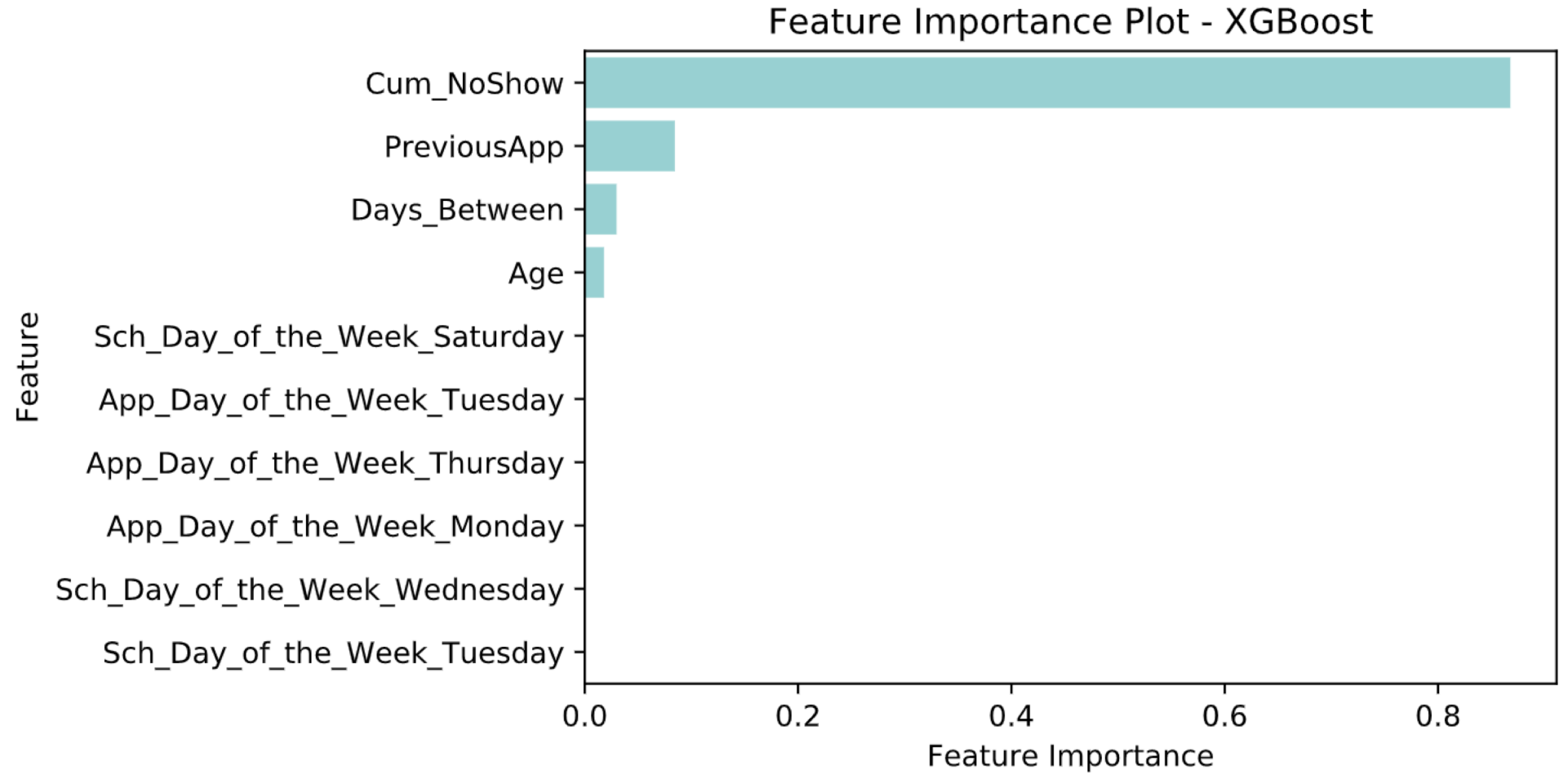
F-1 Score: 0.83
Accuracy: 0.92

ACTUAL VALUES \ PREDICTED VALUES	SHOW	NO-SHOW
SHOW	15,942	1,714
NO-SHOW	40	4,400

Accurately predicted 90% of Shows

Accurately predicted 99% of No-Shows

FEATURE IMPORTANCE



RECOMMENDATION

- Fee Policy
- Automated reminder system, two-way communication
- Extra reminders for those that are predicted to not show up
- Waitlist

FUTURE WORK

- Longer period of Time
- Neighborhood demographics (e.g. median income)
- Accessibility to public transportation
- Effectiveness of reminder method

THANK YOU! QUESTIONS?

APPENDIX

XGBoost	
TN	15773
FN	34
FP	1806
TP	4483
Sensitivity	0.99247288
Precision	0.71283193
Specificity	0.89726378
Balanced Accuracy	0.94486833

Logistic Regression	
TN	15903
FN	131
FP	1676
TP	4386
Sensitivity =	0.97099845
Precision	0.72352359
Specificity	0.90465897
Balanced Acc	0.93782871

Random Forest	
TN	16575
FN	461
FP	1004
TP	4056
Sensitivity =	0.89794111
Precision	0.80158103
Specificity	0.9428864
Balanced Acc	0.92041375

Decision Tree	
TN	16719
FN	812
FP	860
TP	3705
Sensitivity =	0.82023467
Precision	0.81161008
Specificity	0.95107799
Balanced Acc	0.88565633

KNN	
TN	16683
FN	997
FP	896
TP	3520
Sensitivity =	0.77927828
Precision	0.79710145
Specificity	0.94903009
Balanced Acc	0.86415419

SVM	
TN	15681
FN	42
FP	1898
TP	4475
Sensitivity =	0.99070179
Precision	0.70218108
Specificity	0.89203026
Balanced Acc	0.94136603

SVM GridSearchCV	
TN	16585
FN	1048
FP	994
TP	3469
Sensitivity =	0.7679876
Precision	0.77727986
Specificity	0.94345526
Balanced Acc	0.85572143