Predicting Patient No Shows with Machine Learning

Chase Kregor
INFO 4604 Applied Machine Learning Final Project
https://github.com/chasekregor/NoShows

Goals

Can we predict when a patient is or isn't going to show up for an appointment based on various patient attributes and scheduling?

Context: Having less no shows in hospitals means more revenue, higher employee satisfaction and efficiency, and most importantly higher levels of care for patients.

The Dataset: “Medical Appointment No Shows” on Kaggle

- https://www.kaggle.com/joniarroba/noshowappointments
- 300k medical appointments of the public healthcare in the capital city of Espirito Santo State - Vitoria - Brazil
- 15 different patient characteristics and if they showed up or not

The Features

Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension	Diabetes	Alcoholism	Handcap	SMS_received	No-show
F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1	0	0	0	0	No
M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	0	0	0	0	No
F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	0	0	0	0	No
F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	0	0	0	0	No
F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1	1	0	0	0	No

Preprocessing:

- Making sure nothing was blank.
- Got rid of negative age values
- Turned features into binary attributes
- Calculating waiting time proved to be quite difficult

The Algorithms I Turned Into An Ensemble

Logistic Regression

- Has proven to work well on medical data in the past
- Easily comprehensible
- Familiarity

Random Forest

- Decision Trees
  - Could easily explain this to a doctor i.e “if this patient is an alcoholic the likelihood they miss an appointment goes up”
- Was a bit worried about overfitting so I limited the depth of the tree

K-nearest neighbors

- Base on majority vote in features
  - Figured this would be good for “public health”
  - Not necessarily precision.

Error Analysis

Confusion Matrix

```
In [207]: from sklearn.metrics import confusion_matrix

          confusion_matrix(testing_labels, ensemblepredictions)

Out[207]: array([[17658,    11],
                 [ 4429,     8]])
```

- Only able to reach 80% accuracy, I was unhappy with this.
- This is a complex problem
- There isn't enough discrepancy in the features to simply predict whether someone is simply going to show or not show up.
- That being said we can predict that certain people are more likely to show up than other though we kind of already had these intuitions.
    - Example: Age: Older people show up more than younger people
- Possibly with more data or features we might be able to predict if people are or aren't going to show up in Brazil.
- It might just be this individual dataset, maybe you could use the same models with the same features but different and or more robust data and it might work.

Conclusion

Goal: Can we predict when a patient is or isn't going to show up for an appointment based on various patient attributes and scheduling?

Result: In all honesty, maybe? At least I wasn't able to better than human intuition. Maybe with a different dataset and or more features we might be able to?

Result: Takeaway, the world is very complex and random. It is isn't always easy to predict things, even with machine learning.

Any Questions?

Thanks for listening. Check out my github for the repository.