Driven Data: Blood Donor Predictions

What is the Data Science Question?

Given an individuals blood donation history, can we predict whether s/he donated during a blood drive in March 2007?

```
training_data.head()
```

	ID	Last_Donation	Number_of_Donations	Volume_Donated	First_Donation	March_Donation
0	619	2	50	12500	98	1
1	664	0	13	3250	28	1
2	441	1	16	4000	35	1
3	160	2	20	5000	45	1
4	358	1	24	6000	77	0

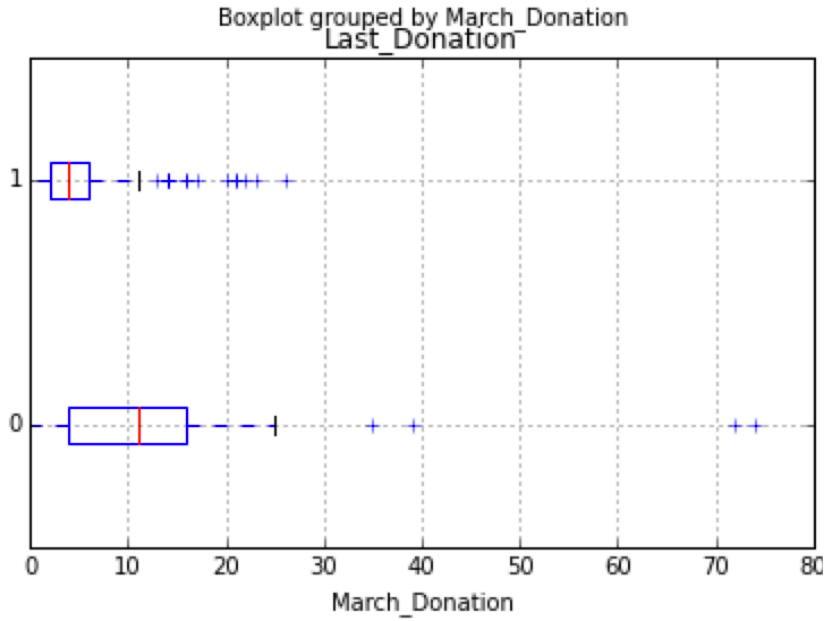
Reuben Karchem

2. What's the data?

Data Source: UCI Machine Learning Repository

Taiwanese mobile blood transfusion center

Training and Test data sets provided



Existing Features:

1) months since last donation
2) total number of donation
3) total blood donated (cc)
4) months since first donation.

Outcome:

Did the individual donate blood in March 2007?

Binary: Yes = 1, No = 0

What new, informative features can be derived from existing features? Donation interval? Non-linear features?

Do you think the data will enable to develop a effective model? Donor data would be tremendously useful

3. What's next?

Is there actually a difference between donors and non-donors across these features?

- Descriptive statistics using pd.groupby
- Corresponding box plots

Initial models

- KNN – neighbors = 5
- Logistic regression

```
#KNN
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train, y_train)
y_pred2 = knn.predict(X_test)
print metrics.accuracy_score(y_test, y_pred2)
```

Cross-validation for KNN

What other evaluation metrics would be informative?

Can new features be introduced?

```
#LOGISTIC REGRESSION
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression(C=1e9)
from sklearn.cross_validation import train_test_split
feature_cols = ['First_Donation', 'Last_Donation','Number_of_Donations','Volume_Donated']
X = training_data[feature_cols]
y = training_data.March_Donation
X_train,X_test, y_train, y_test = train_test_split(X, y, random_state=3)
```