Study of dataset on Diabetes from 130 US hospitals over 10 years

- Chang Zhou
- Harsha Jakkappanavar
- Savan Patel
- Jenny Johns

Introduction / Background

Random Forest

Conclusion

Exploratory Analysis

Logistic Regression

FP-Growth

Introduction / Background

- Consists of 100,000 records of patients diagnosed with diabetes collected from 130 US hospitals over 10 years.
- Some of the questions we intend to answer from this dataset is whether we could predict readmission of a patient given certain attributes and whether there are any associations between the diabetic medications given to patients.

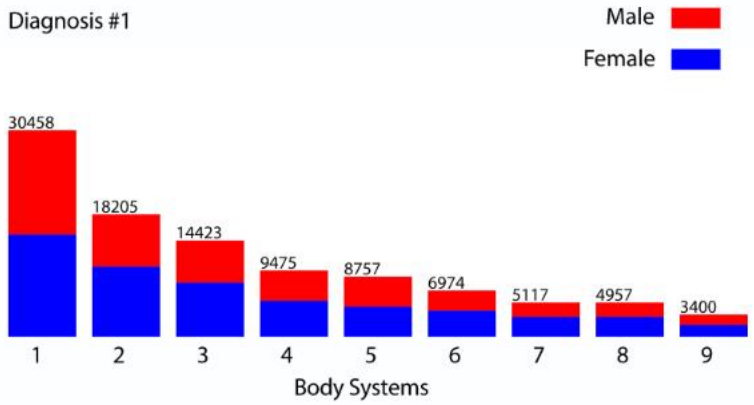
Exploratory Analysis

- There are 101, 766 patient encounters
- 55 features for each encounter

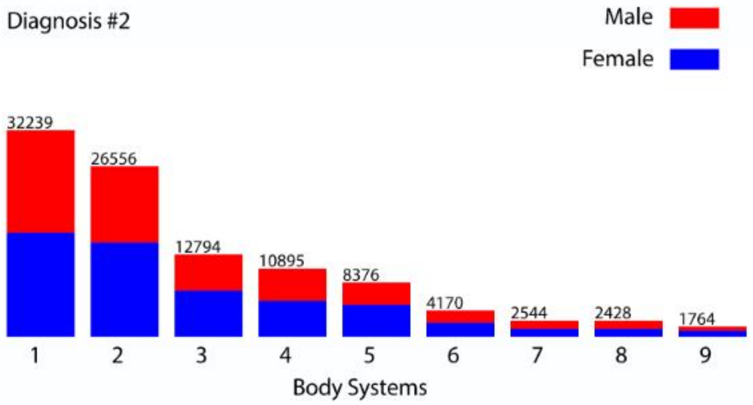
ICD9 values to Body systems

Preprocessing

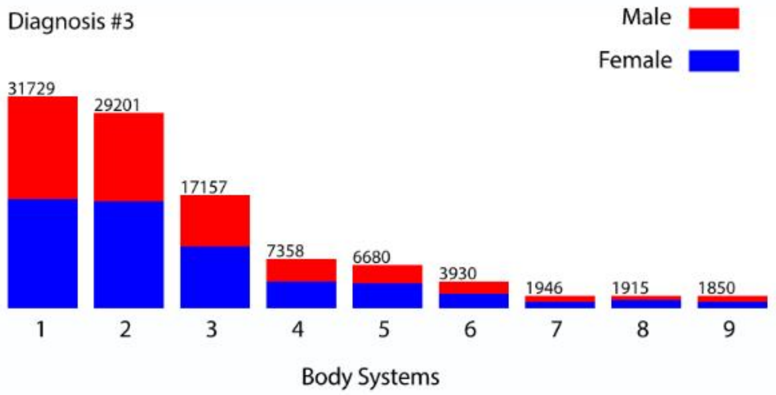
ICD9 values to Body systems



1. Circulatory
2. Other
3. Respiratory
4. Digestive
5. Diabetes
6. Injury
7. Genitourinary
8. Musculoskeletal
9. Neoplasms



1. Circulatory
2. Other
3. Diabetes
4. Respiratory
5. Genitourinary
6. Digestive
7. Neoplasms
8. Injury
9. Musculoskeletal



1. Circulatory
2. Other
3. Diabetes
4. Respiratory
5. Genitourinary
6. Digestive
7. Injury
8. Musculoskeletal
9. Neoplasms

Preprocessing

As part of preprocessing we identified that the following features were unuseful and hence dropped them.

- Encounter Id
- Patient number
- Weight
- Payer code
- Medical specialty

FP-Growth

Attributes:

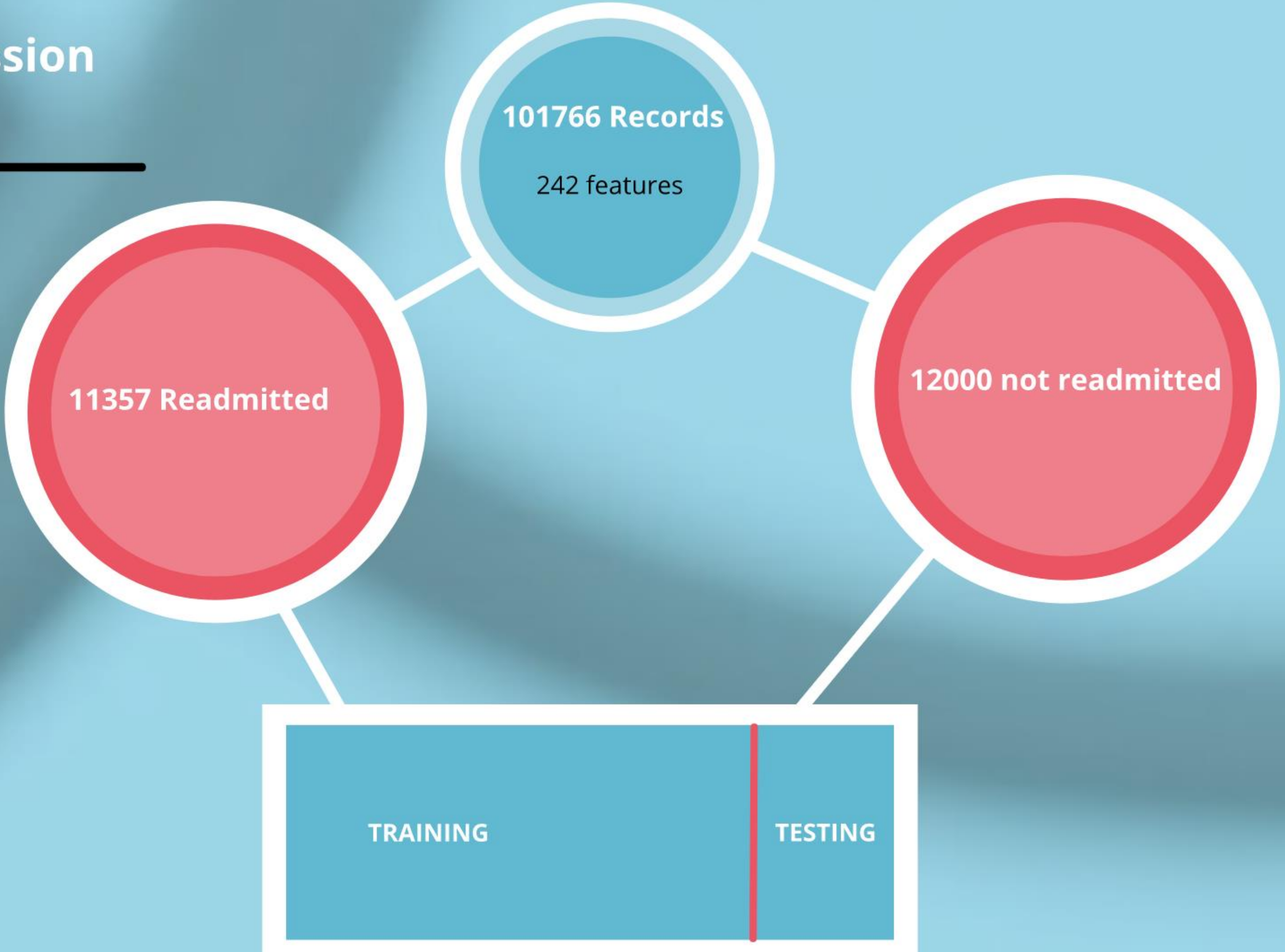
- metformin
- repaglinide
- nateglinide
- chlorpropamide
- glimepiride
- acetohexamide
- glipizide
- glyburide
- tolbutamide
- pioglitazone
- rosiglitazone
- acarbose
- miglitol
- troglitazone
- tolazamide
- examide
- citoglipton
- insulin

We found 2 rules:

metformin ==> insulin with Confidence 0.5 and Lift 0.72

insulin ==> metformin with Confidence 0.18 and Lift 0.72

Logistic Regression



101766 Records

242 features

11357 Readmitted

12000 not readmitted

TRAINING

TESTING

Results

Training Data: 242 features
Model Accuracy: 56.57%

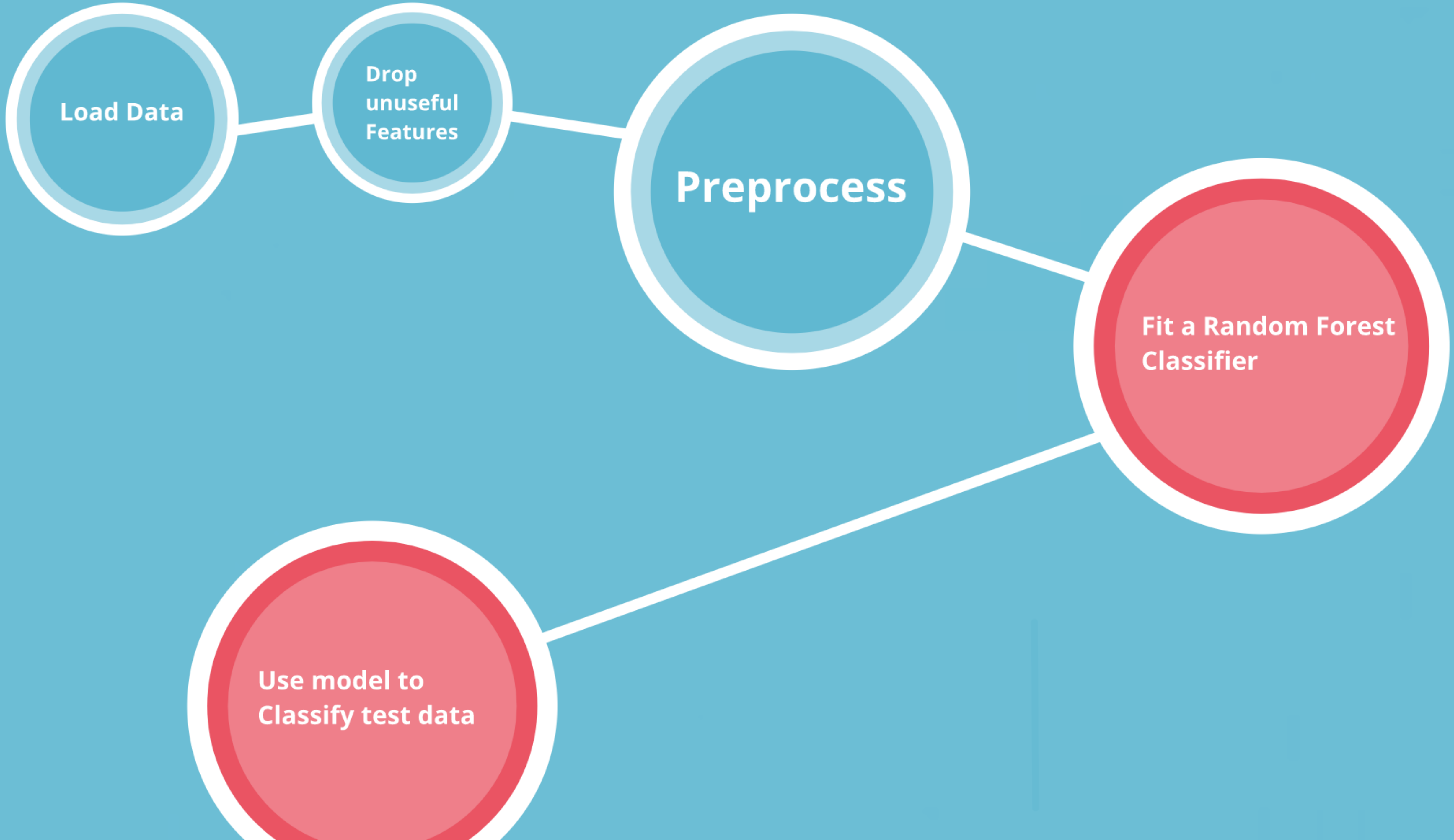
Applied PCA to reduce dimensionality to 6 (48% variance in original data)
Model Accuracy: 60.64%

Random Forest

- Ensemble learning method for classification
- It constructs a large number of decision trees on the training data to create a classifier model, this model is later used on the test data for prediction.

Preprocess

Results



Load Data

Drop unuseful Features

Preprocess

Fit a Random Forest Classifier

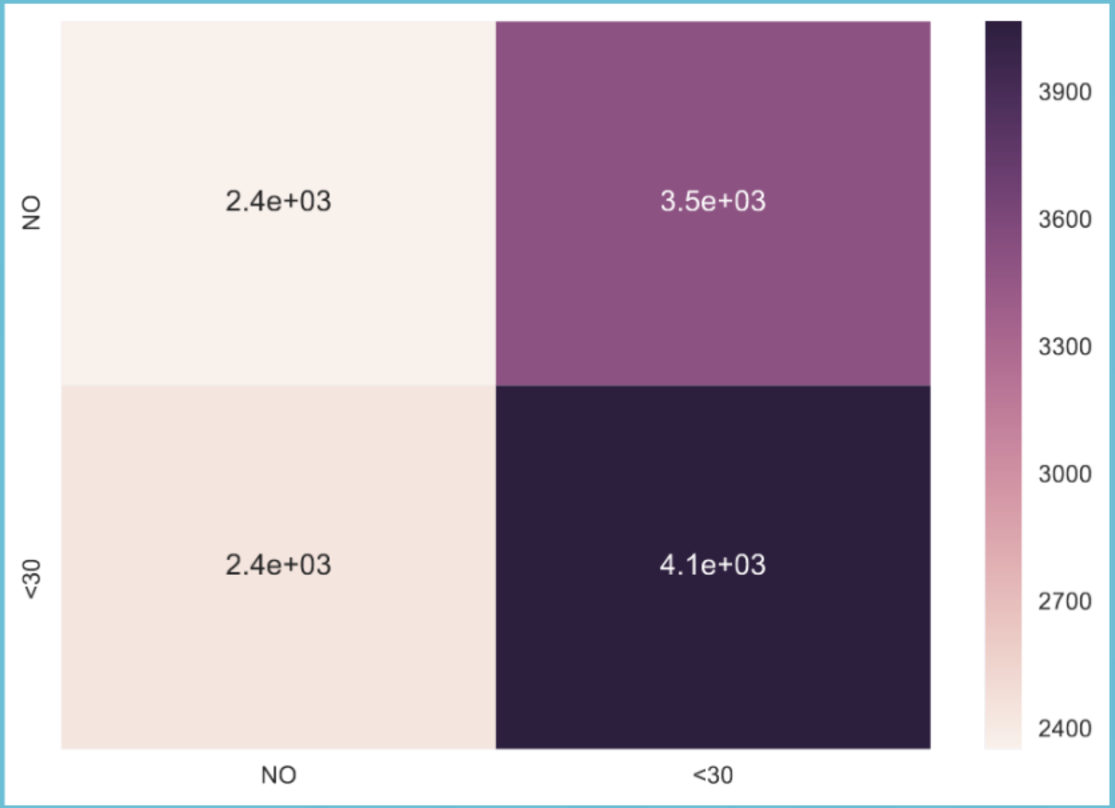
Use model to Classify test data

Results

Training Data: 44 features

Model Accuracy: 51.19%

Confusion Matrix



Conclusion

Unsupervised

- FP-Growth
- PCA

Supervised

- Logistic Regression
- Random Forests