Predicting Rain in Sydney, Australia

Kristofer Schobert

Our Goal

Use a supervised learning classifier to answer:

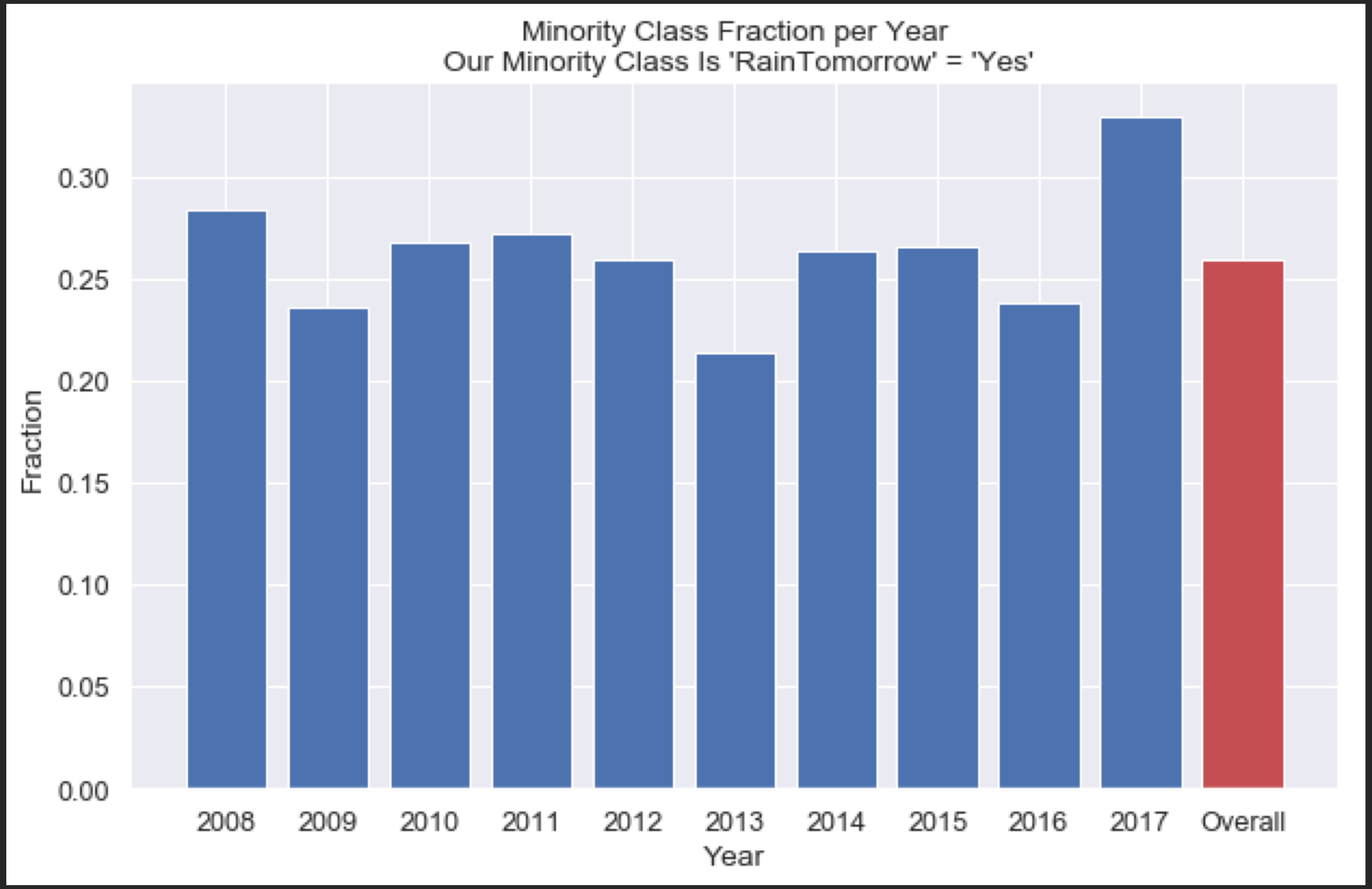
“Will it rain tomorrow in Sydney, Australia?”

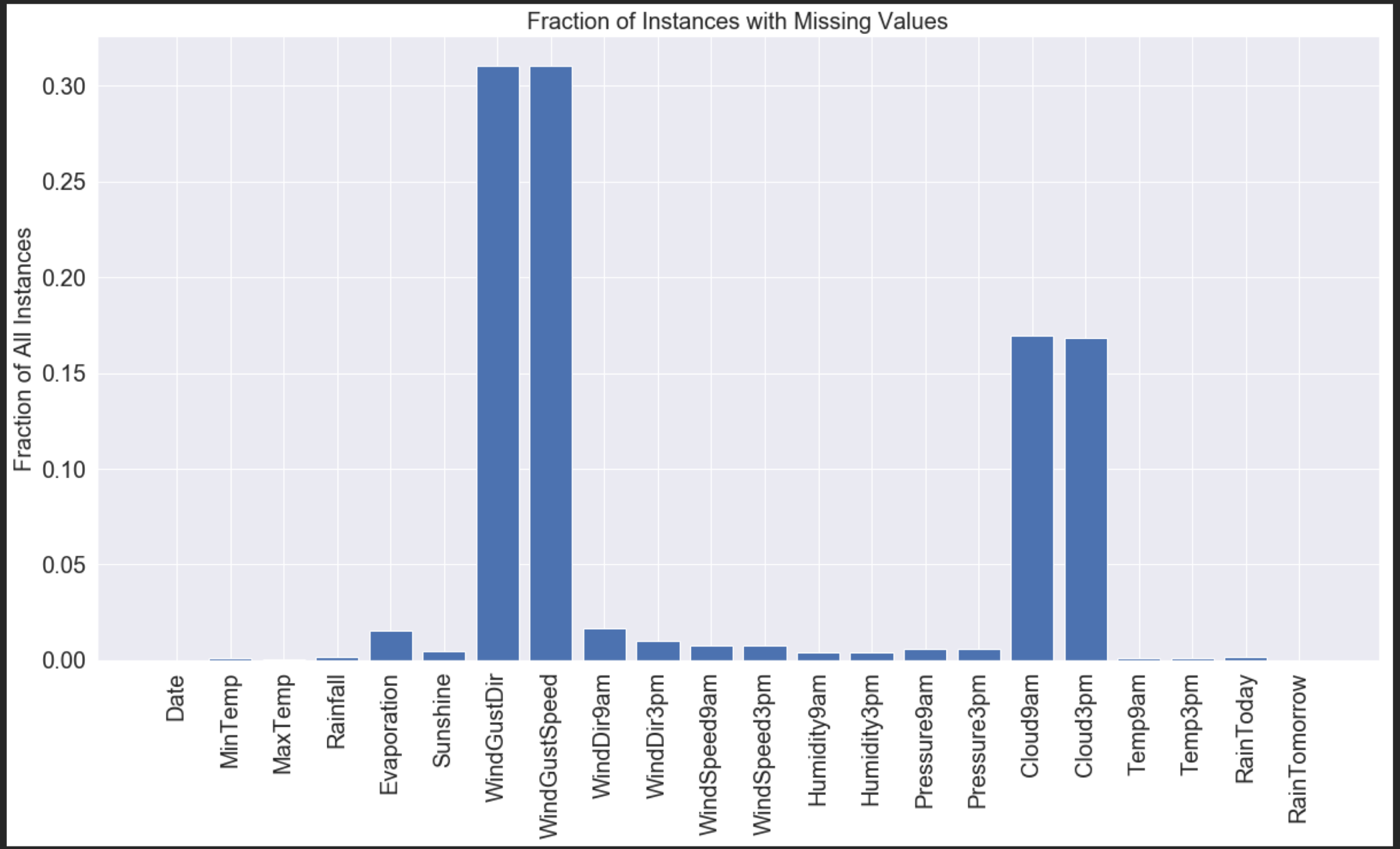
Significance of Weather Predictions

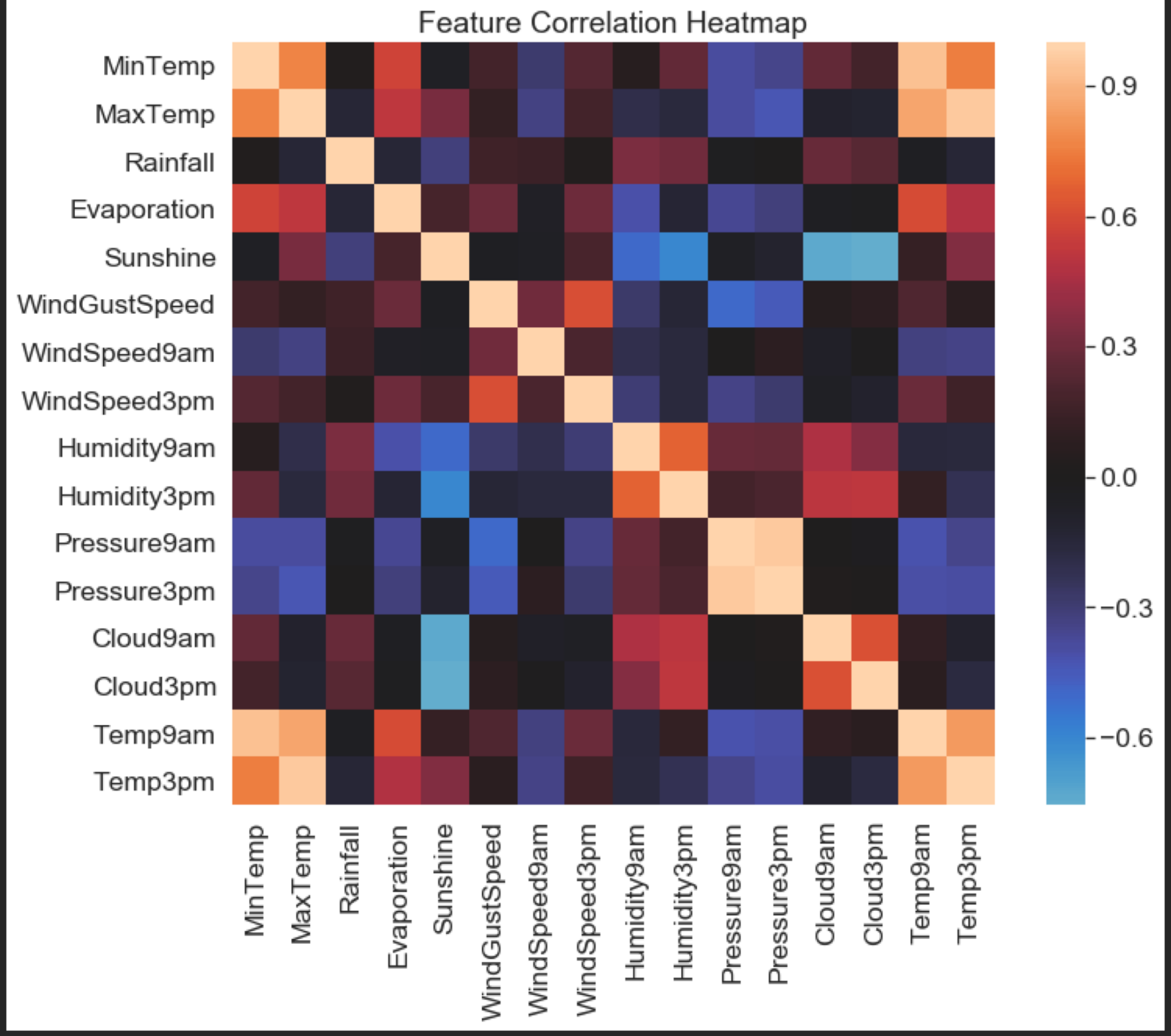
- Farmers
- Transit
- Safety
- Future of our planet

Our Dataset

- Daily Weather Observations
  - Several Stations
  - Several Australian Cities
  - 2007 - 2017
- Features
  - Twenty-two measurements of meteorology
- Outcome Variable
  - Rain Tomorrow

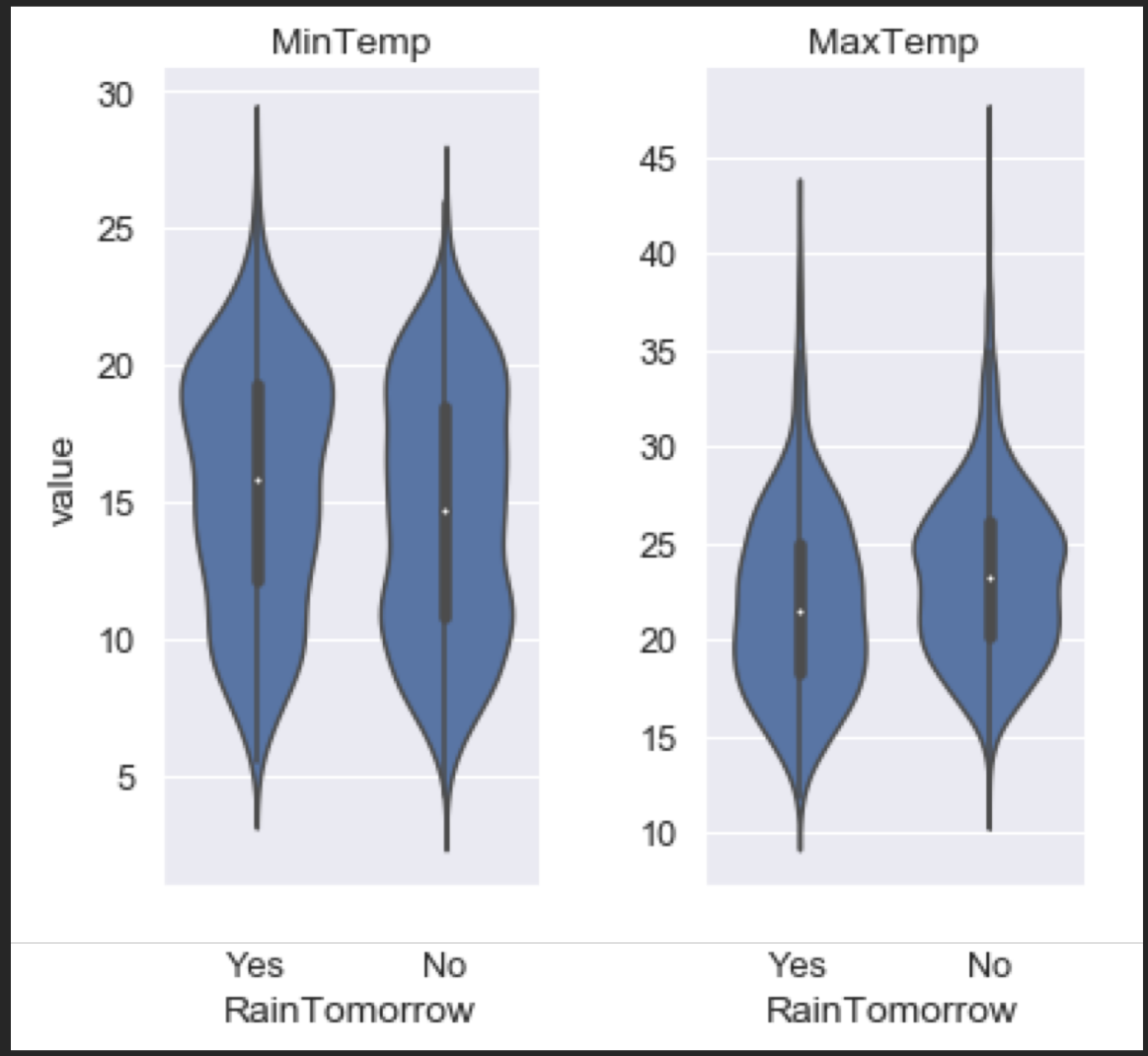

Minority Class Fraction per Year
Our Minority Class Is 'RainTomorrow' = 'Yes'
Fraction
0.00
0.05
0.10
0.15
0.20
0.25
0.30
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Overall
Year


Fraction of Instances with Missing Values
Fraction of All Instances
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Date
MinTemp
MaxTemp
Rainfall
Evaporation
Sunshine
WindGustDir
WindGustSpeed
WindDir9am
WindDir3pm
WindSpeed9am
WindSpeed3pm
Humidity9am
Humidity3pm
Pressure9am
Pressure3pm
Cloud9am
Cloud3pm
Temp9am
Temp3pm
RainToday
RainTomorrow


Feature Correlation Heatmap
MinTemp
MaxTemp
Rainfall
Evaporation
Sunshine
WindGustSpeed
WindSpeed9am
WindSpeed3pm
Humidity9am
Humidity3pm
Pressure9am
Pressure3pm
Cloud9am
Cloud3pm
Temp9am
Temp3pm
0.9
0.6
0.3
0.0
−0.3
−0.6

Feature Creation

- TempRange
- Month



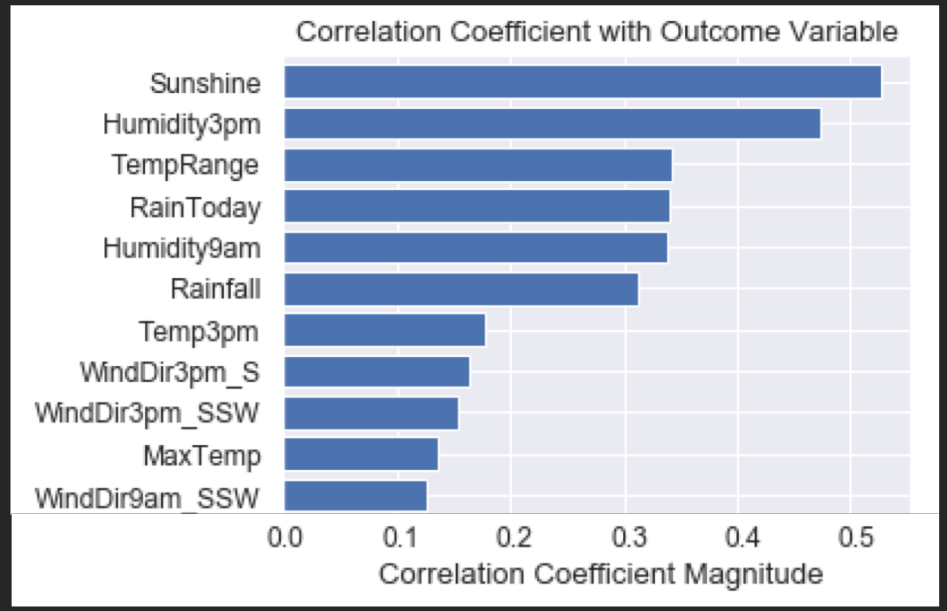
Method for Model Evaluation

- Classifiers
  - Nearest Neighbors
  - Random Forest
  - Logistic Regression
  - Support Vector Machine
  - Gradient Boosting

Method for Model Evaluation

- Vary model parameters

- Vary datasets
  - df_all
  - df_sign
  - df_pca
    - 5 components



Method for Model Evaluation

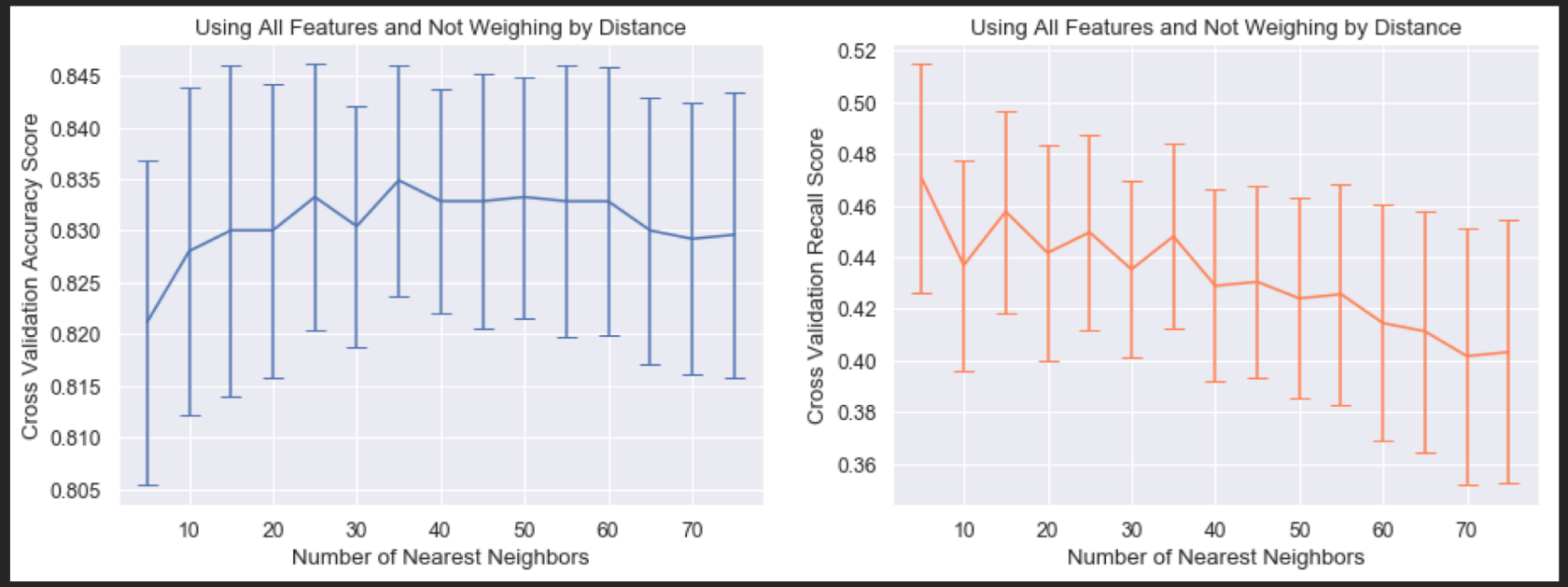
- Winning Criterion
  - Best Mean Cross Validation **Accuracy** Score
  - If tie: Best Mean Cross Validation **Recall** Score

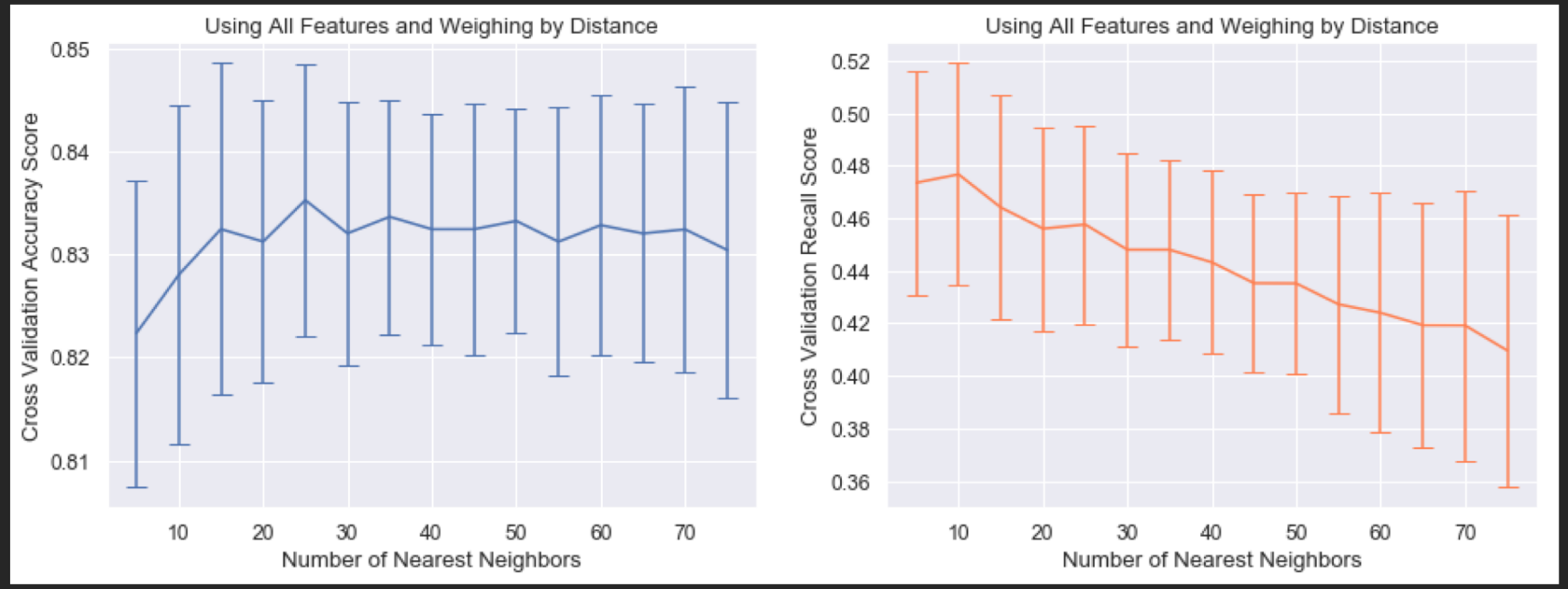
Comparing Model Parameters

How many neighbors?

Weigh by distance or not?

KNN with df_all





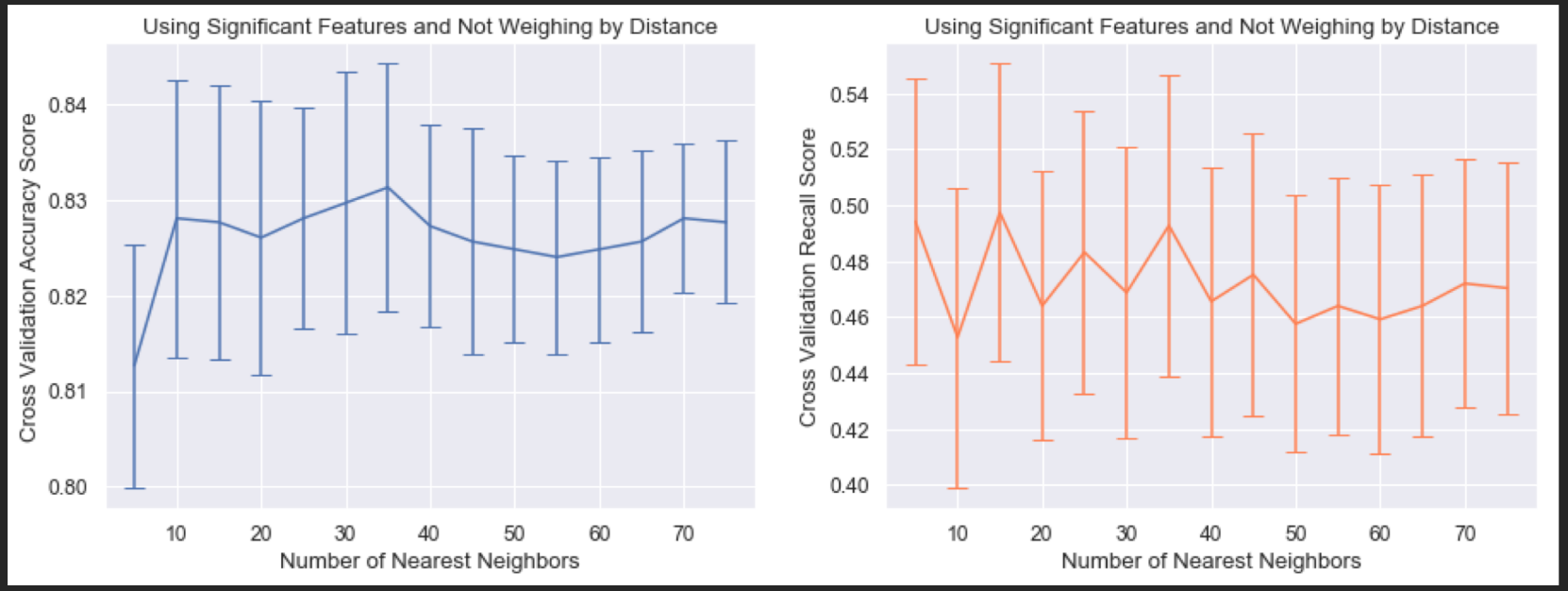
Comparing Model Parameters

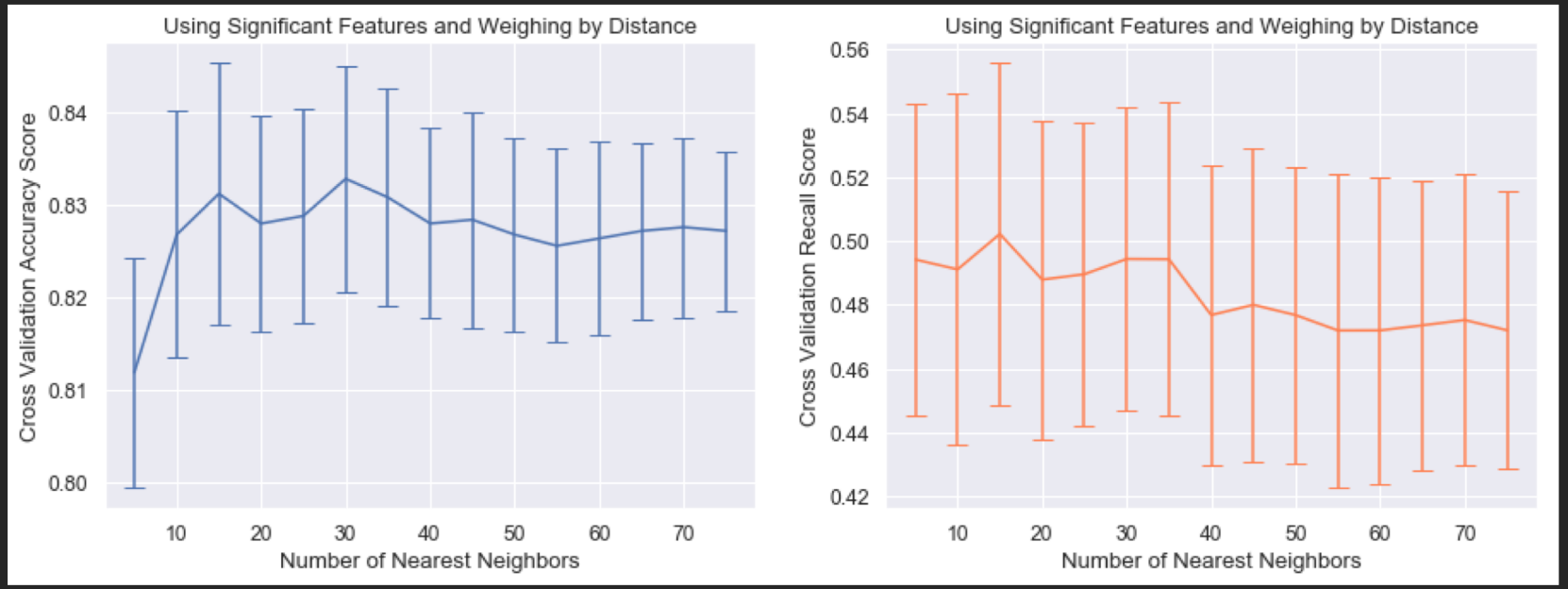
Current winner:

- df_all
- 25 neighbors
- with weights

- Accuracy = 0.835
- Recall = 0.46

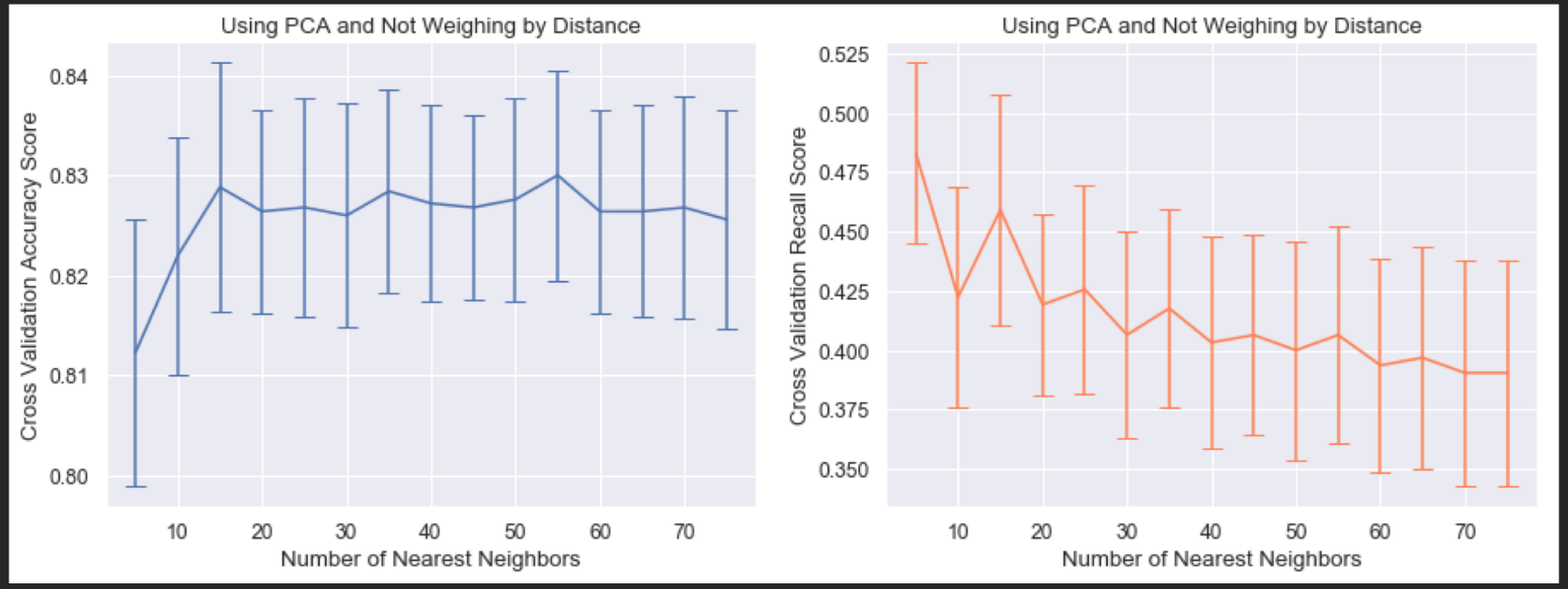
KNN with df_sign

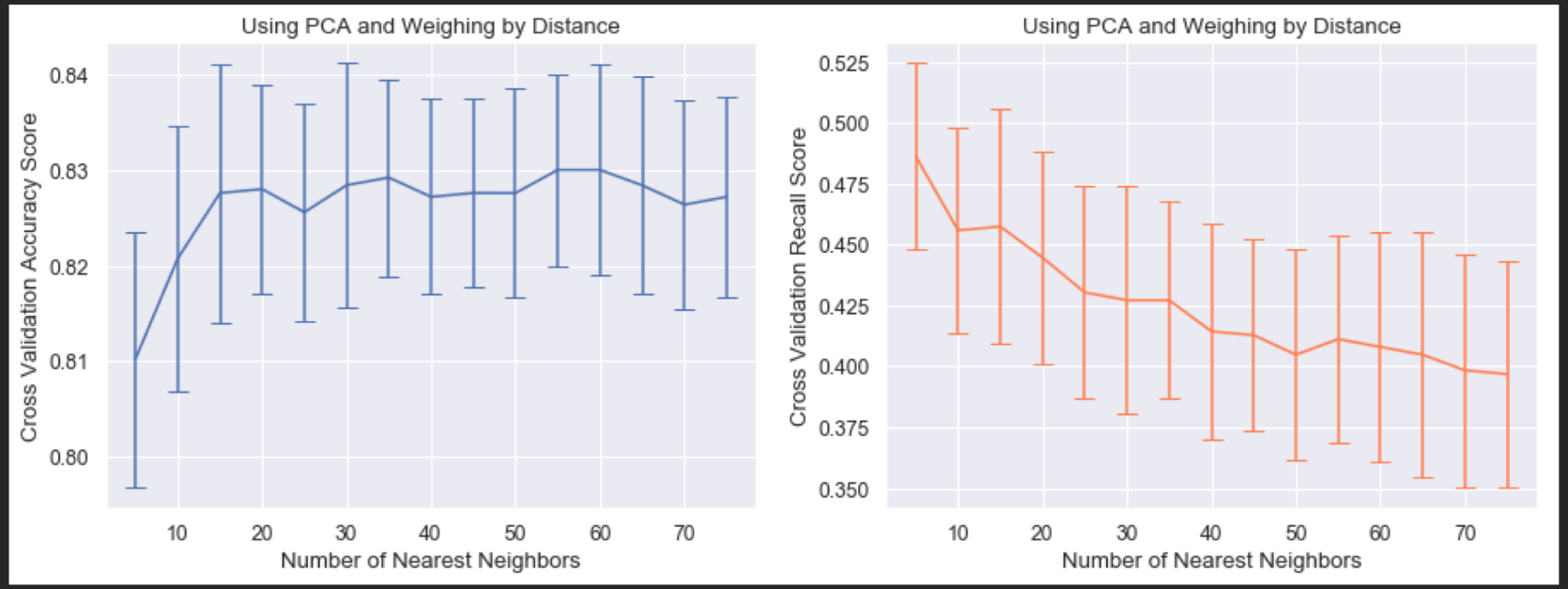




Comparing Model Parameters

KNN with df_pca



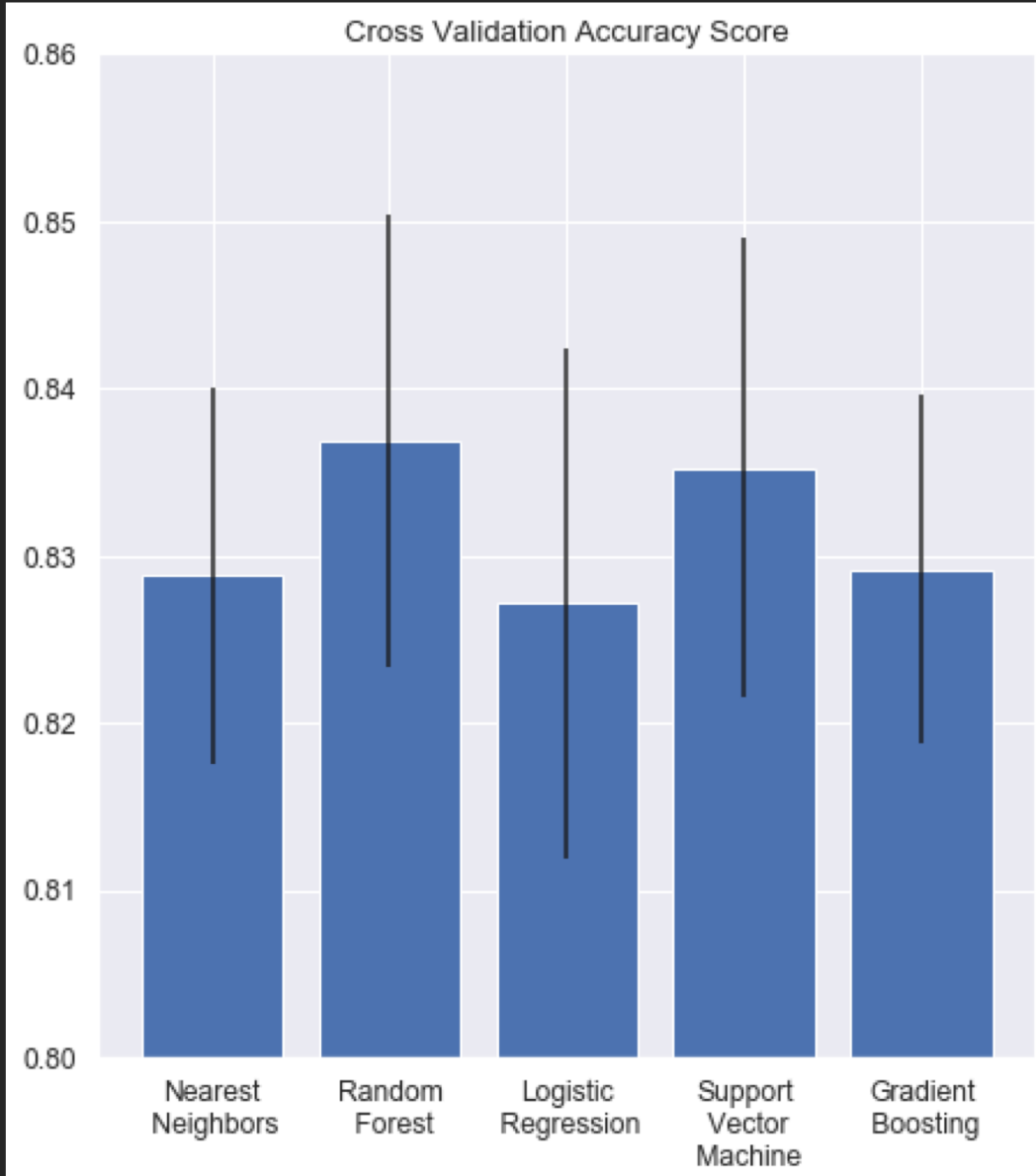


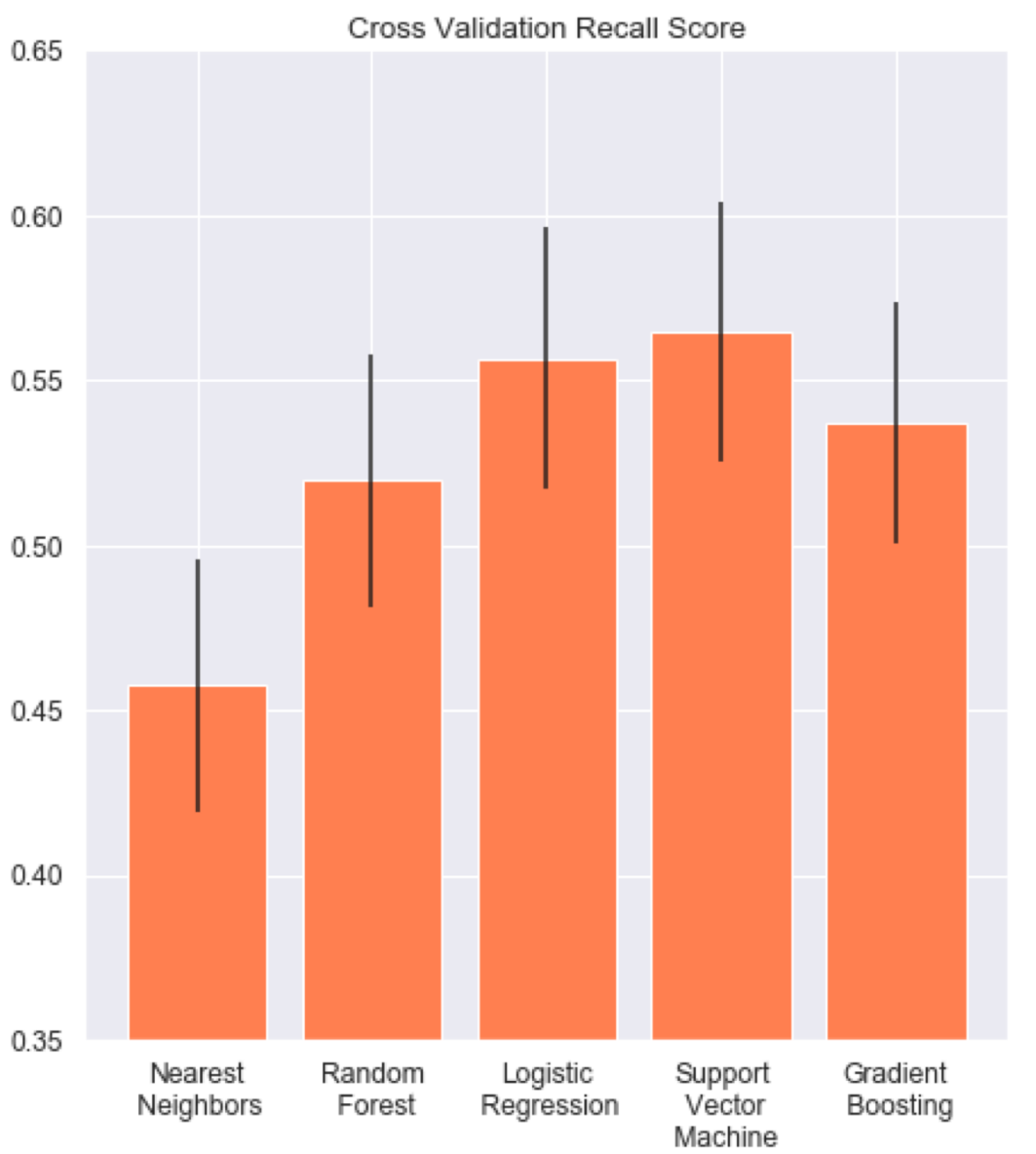
Current winner:

- df_sign
- 30 neighbors
- with weights

- Accuracy = 0.834
- Recall = 0.49

Best of the Five Models





Our Winner:
Support Vector Machine

Prediction of Testing Data (2011 and 2014)

Confusion Matrix

[[462 39]
 [77 100]]

Scores

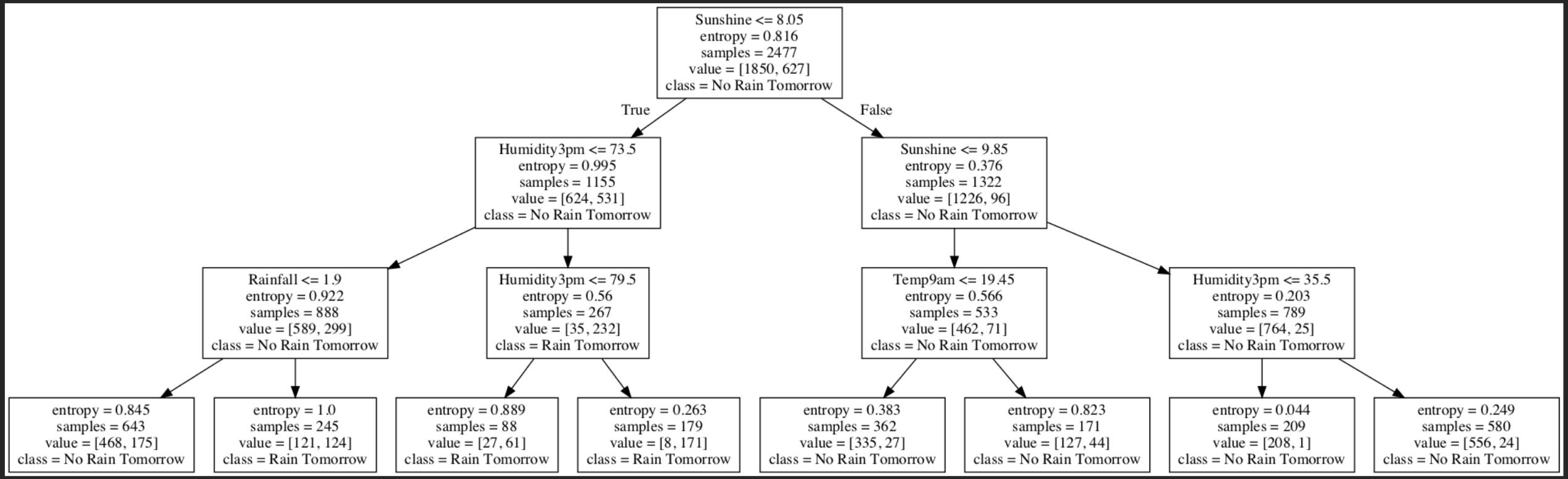
Accuracy: 82.9%

Recall: 56.5%

Specificity: 92.2%

Model Shortcomings – Feature Importance

Decision Tree



Random Forest Feature Importance