Predicting Rain Patterns in Australia

Justin Grisanti

Objectives – 5 Step Model

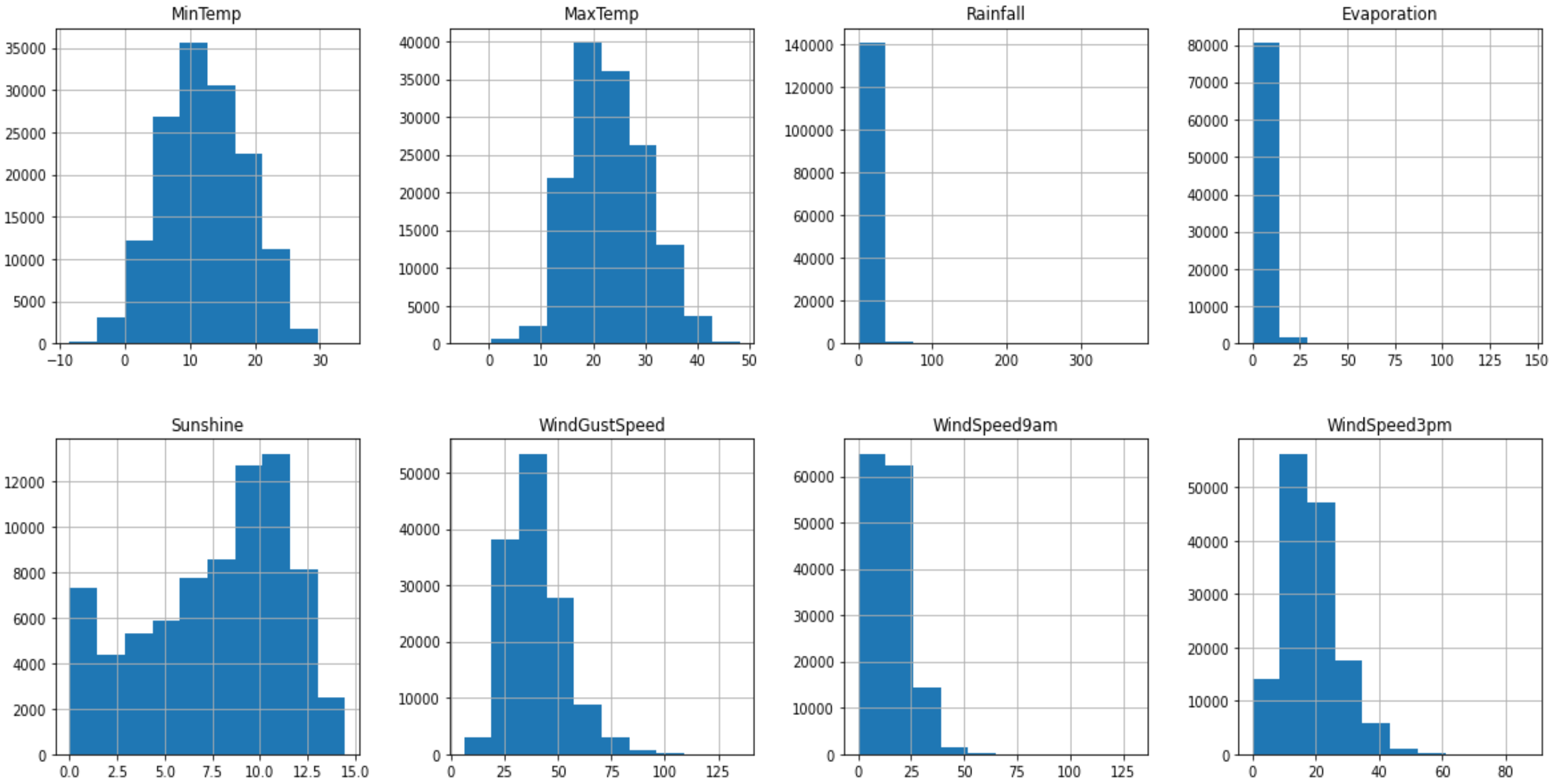
- Obtain a Business Understanding
- Understand the Data
- Prepare Data for Modeling
- Create Models
- Generate Results

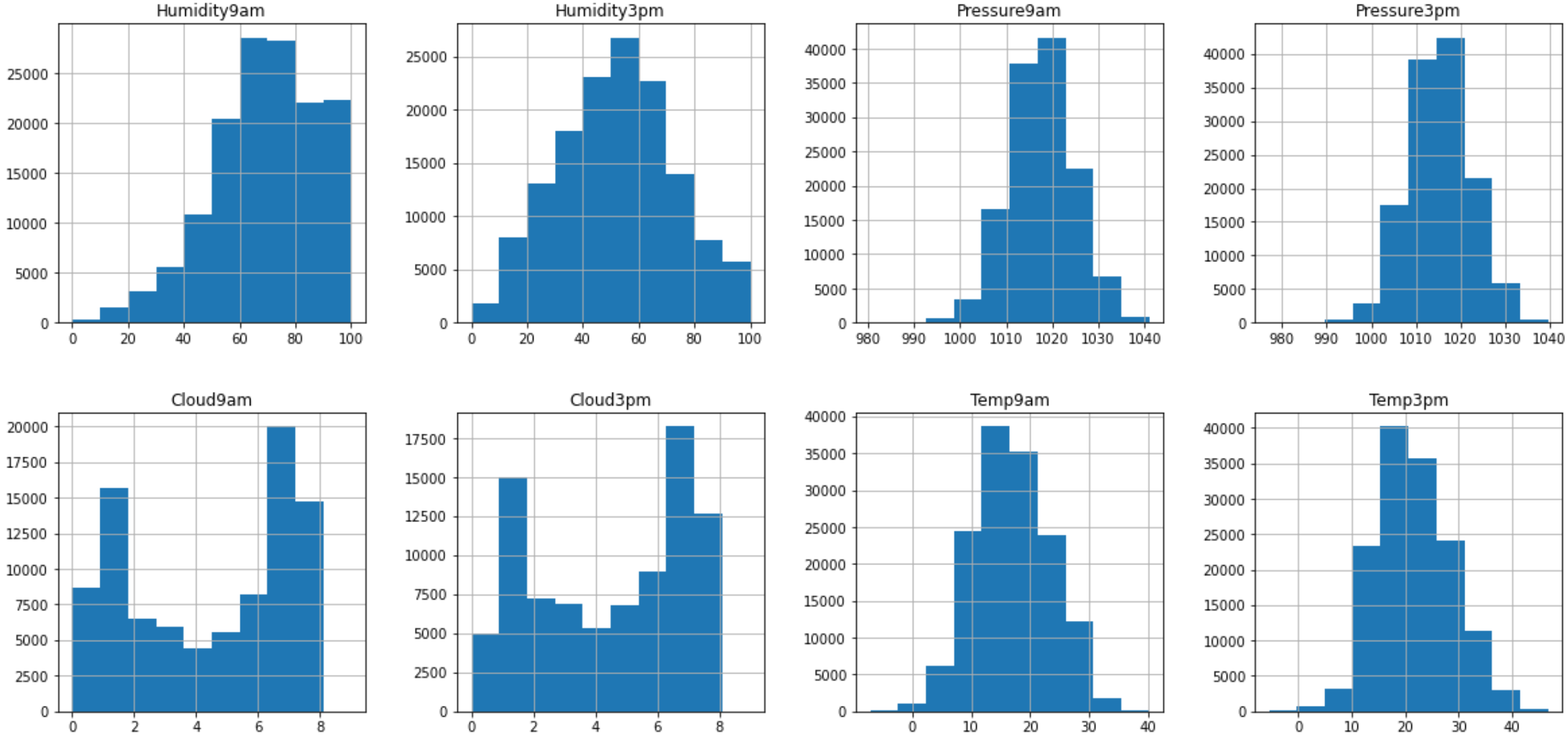
Business Understanding

- Stakeholders:
  - Australian Government
  - Citizens of Australia
  - Firefighters of Australia
- Business Problem:
  - To understand which attributes best predict whether or not it will **rain the next day**
- Predictive Classification Model

Data Understanding

- Various Weather Data
  - Wind Patterns
  - Sunshine Data
  - Cloud Data
  - Humidity
  - Temperature
  - Etc.
- Histograms of Columns


MinTemp
MaxTemp
Rainfall
Evaporation
Sunshine
WindGustSpeed
WindSpeed9am
WindSpeed3pm


Humidity9am
Humidity3pm
Pressure9am
Pressure3pm
Cloud9am
Cloud3pm
Temp9am
Temp3pm

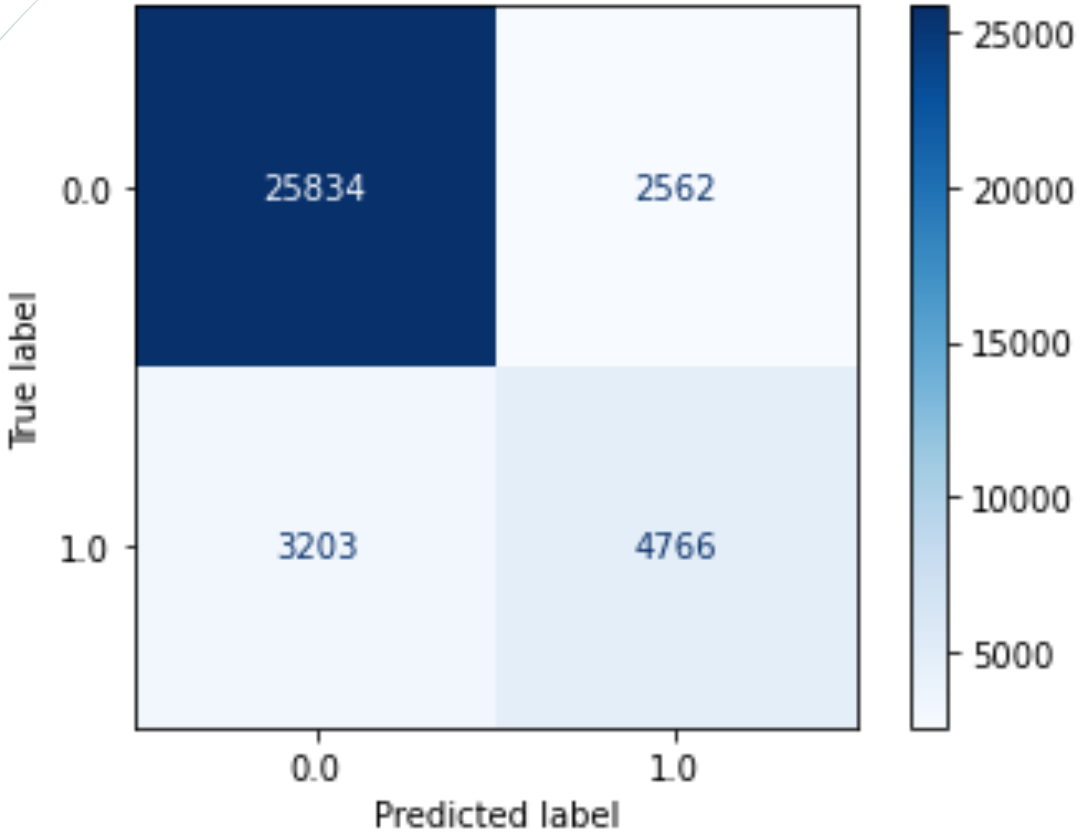
Data Preparation

- Create Dummy Variables
- Train and Test Split:
  - Fill empty data
  - Check for erroneous data
  - Normalize data
- Clean all columns to ensure better modeling

Modeling

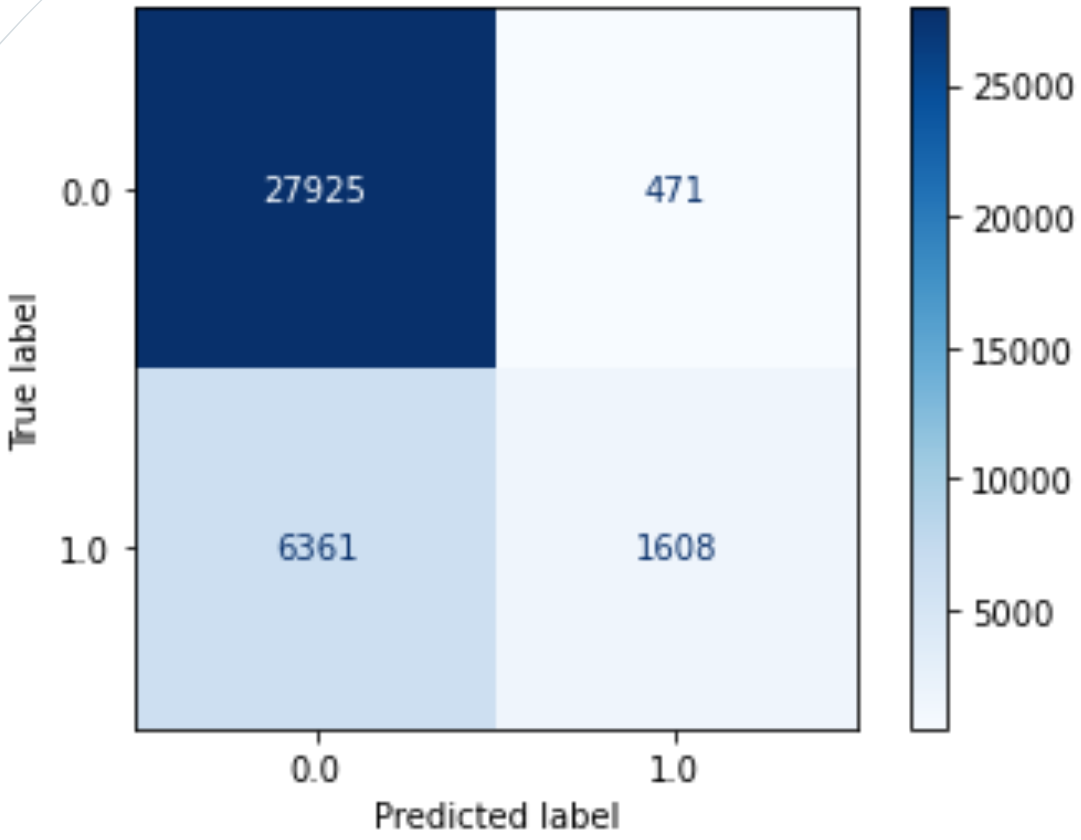
- **Logistic Regression**
  - Best Model: **84.19%** Accurate
- K-Nearest Neighbors
  - Best Model: 81.21% Accurate
- Decision Trees
  - Best Model: 82.44% Accurate

Logistic Regression: Confusion Matrix



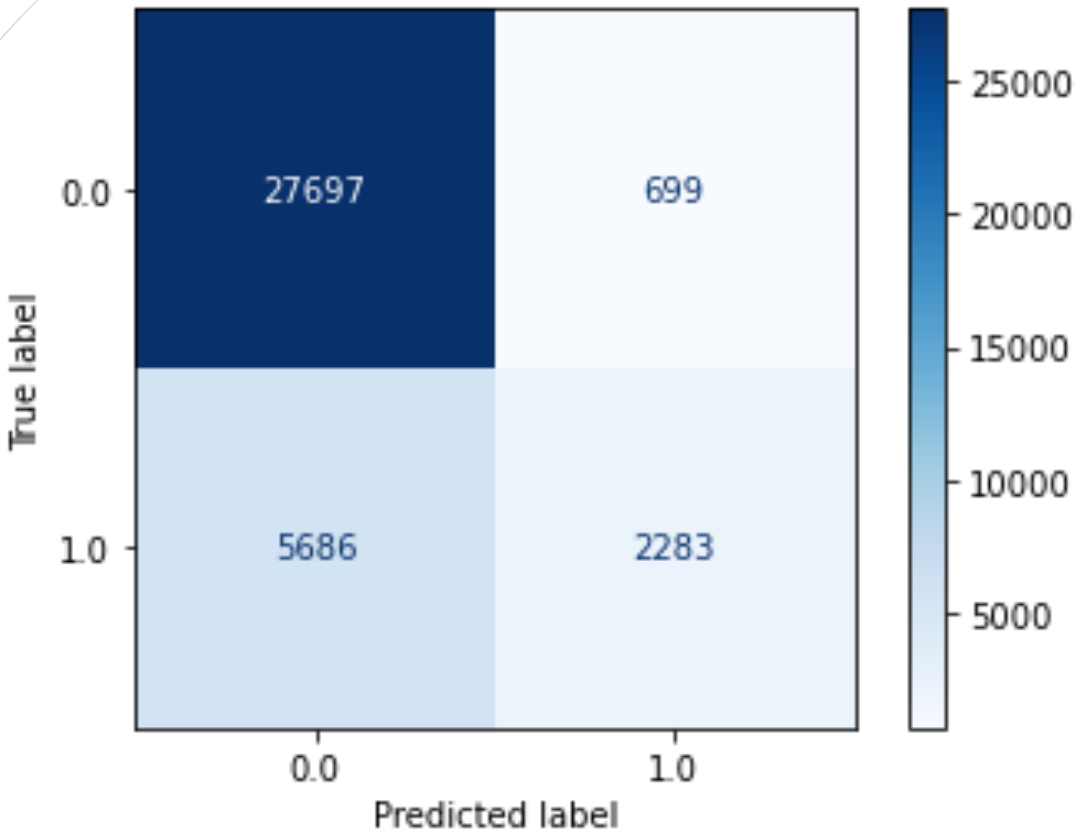
- 84% Accurate
- 7% Type I Error
- 9% Type II Error

K-Nearest Neighbors: Confusion Matrix



- 81% Accurate
- 1% Type I Error
- 18% Type II Error

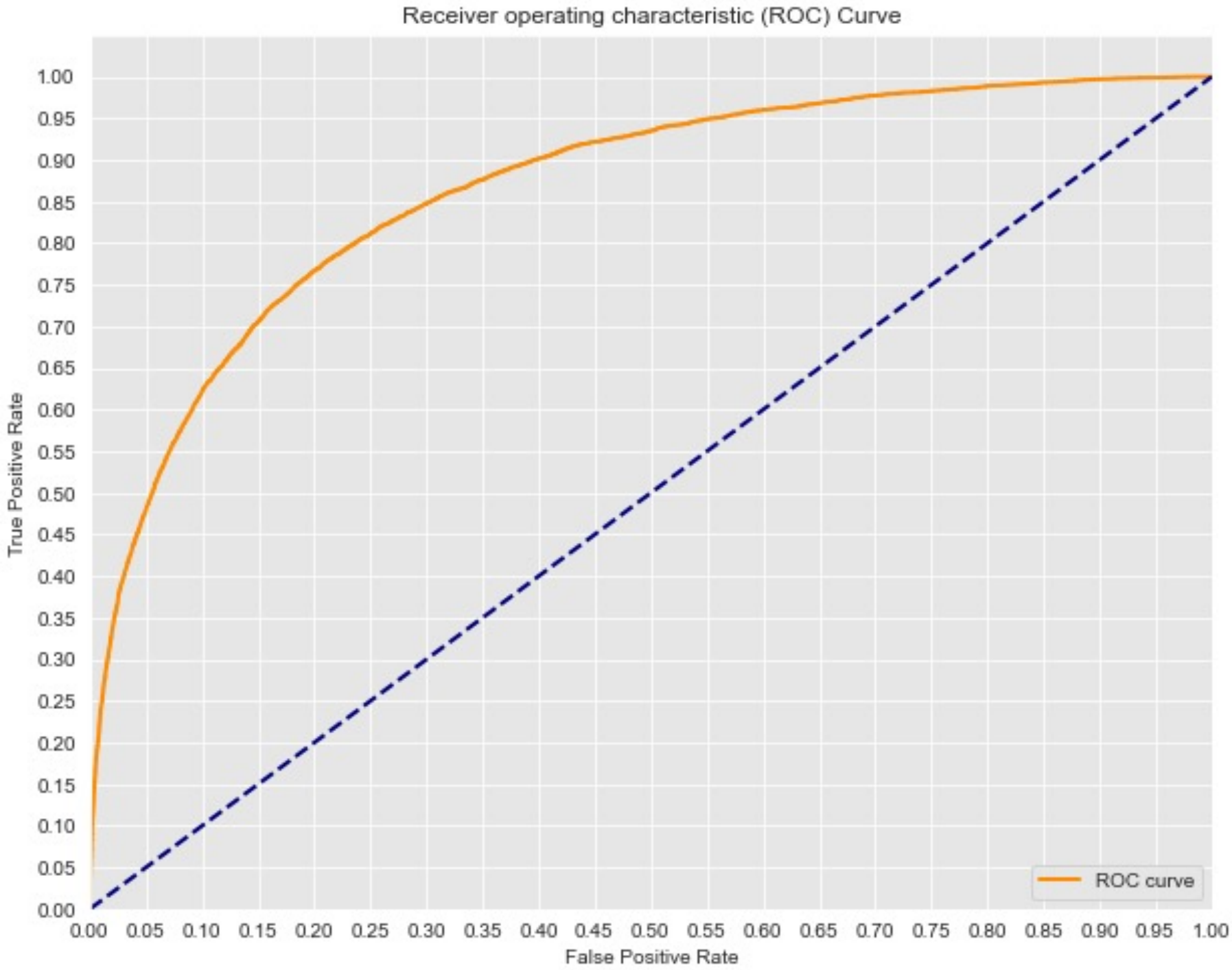
Decision Trees: Confusion Matrix



- 82% Accurate
- 2% Type I Error
- 16% Type II Error

Results

- Best Model: Logistic Regression with an 84% Accuracy Score
  - Precision: 65%
  - Recall: 60%
- AUC: .86
- kNN and Decision Trees have less Type I error but more Type II Error


Receiver operating characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
ROC curve

Next Steps:

- Find more rain data in different periods and see how it performs
- Try a Random Forest Model
- Test more attributes in GridSearchCV

Thank you for your time!