# Forest Cover Prediction Type

By
Abhishek Agarwal
Beau Kramer
Pavan Kurapati

# Overview

The study area includes four wilderness areas located in the Roosevelt National Forest of northern Colorado. These areas represent forests with minimal human-caused disturbances, so that existing forest cover types are more a result of ecological processes rather than forest management practices.

# Understanding the problem

## Labels

Seven Cover Types

1. Spruce Fir
2. Lodgepole Pine
3. Ponderosa Pine
4. Cottonwood/Willow
5. Aspen
6. Douglas-Fir
7. Krummholz

## Features

54 Features

- Elevation
- Aspect
- Slope
- Horizontal and Vertical Distance to hydrology
- Horizontal distance to roadways
- Hillshade 9AM, Noon and 3PM
- Horizontal Distance to Fire Points
- 4 Wilderness Areas
- 40 Soil Types

## Dataset

Training Set  - 15,120

Test Set - 565,892

Attribute Types

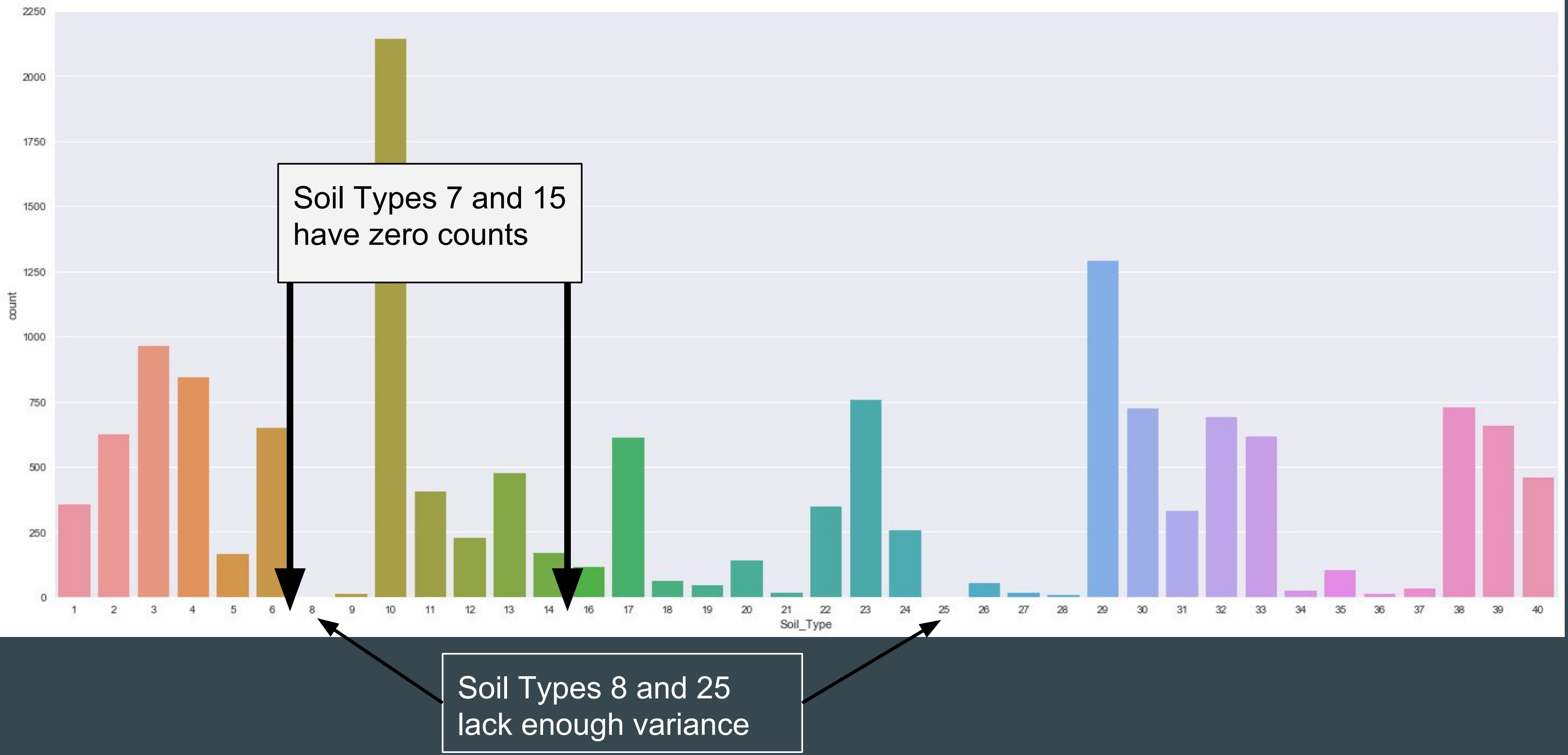
- Numerical - 10
- Categorical - 44
    - 4 Wilderness Types
    - 40 Soil Types

# Project objective:

Predict Tree type for a given 30 x 30 meter cell

# Exploratory Data Analysis

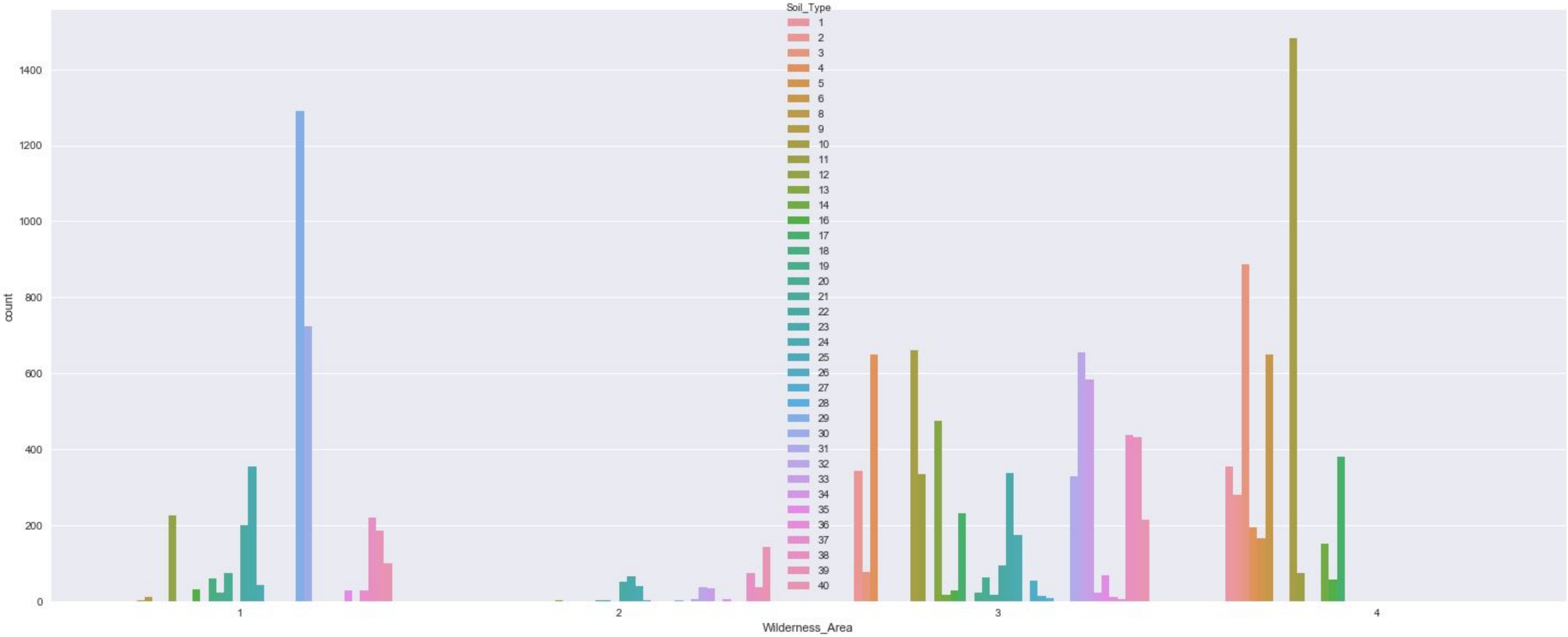
# Soil Counts



Soil Types 7 and 15 have zero counts

Soil Types 8 and 25 lack enough variance

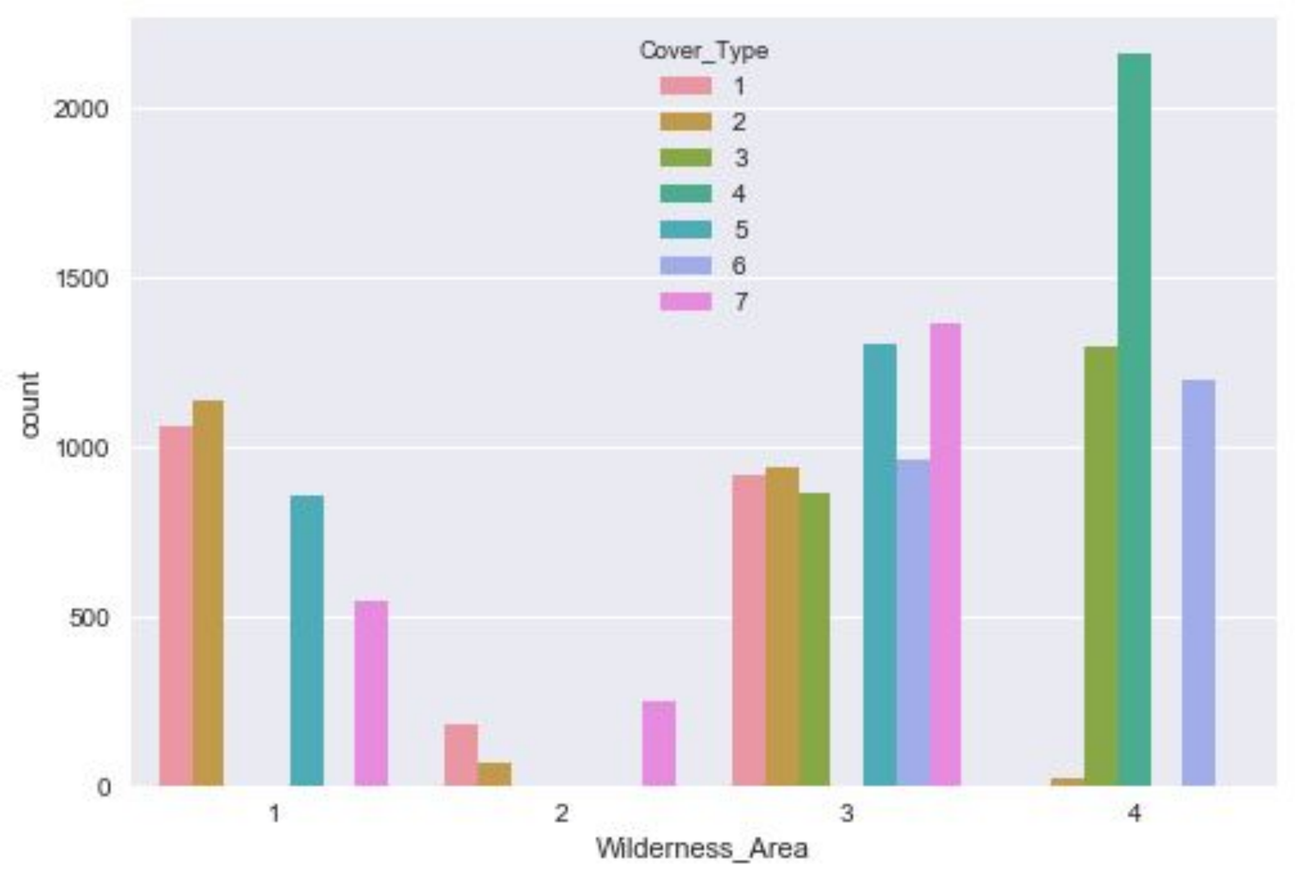
# Wilderness Areas by Soil Type



- Certain Soil Types much more prevalent in specific wilderness areas

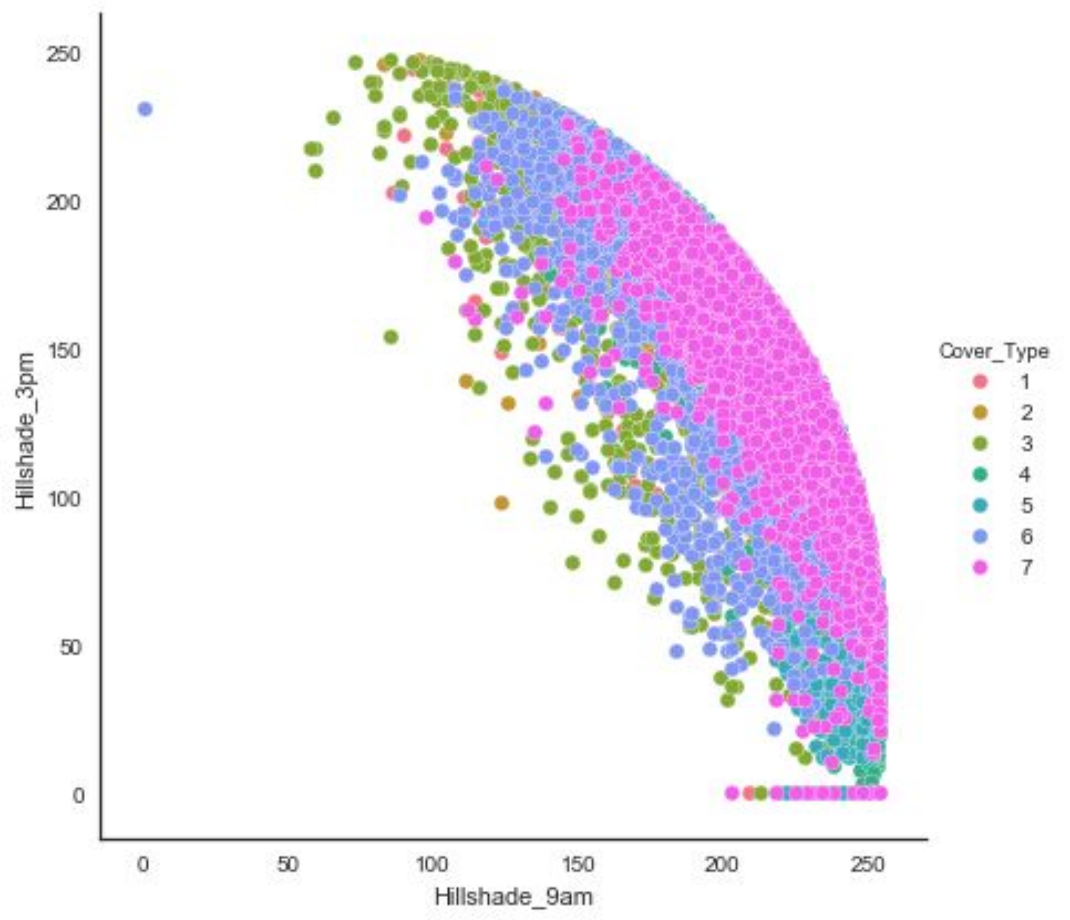
# Wilderness Areas by Tree Cover

- Tree Cover type per Wilderness Area differs significantly
- Eg. - Cover Type 4 is found almost exclusively in Wilderness Area 4

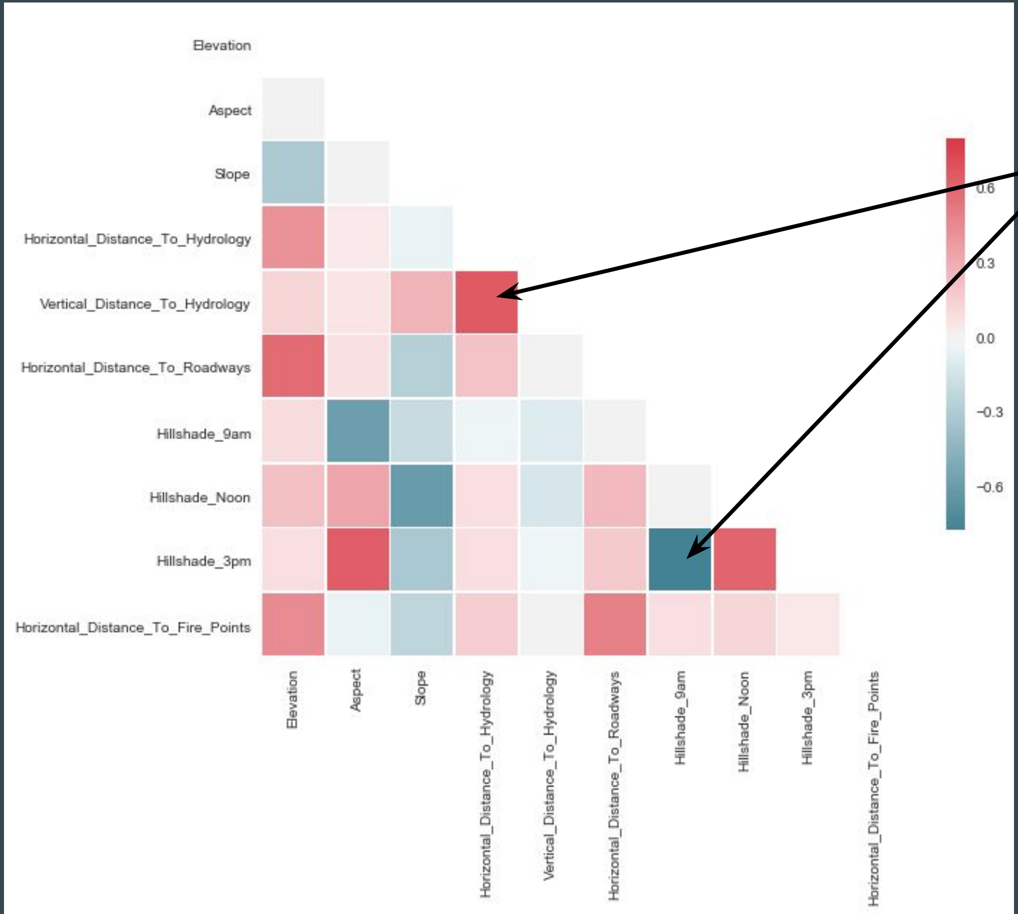


# Hillshade at 9AM and 3PM

- Inverse correlation between the two variables
- Can reduce collinearity by removing one of the variables
- Minimal loss of information



# Correlations



## Highly Correlated

- ➔ Vertical and Horizontal distance to nearest water features
  - ◆ Distance normalized
- ➔ Hillshade Index at 9AM and 3PM
  - ◆ Removed Hillshade at 3PM

# Lesson 1:

## Don’t underestimate data visualization

# Pre-processing

# Feature Engineering

$$Distance\ to\ Hydrology = \sqrt{Horizontal\ distance\ to\ Hyrdology^2 + Vertical\ distance\ to\ Hydrology^2}$$

- Distance to Hydrology variable created to coalesce the Horizontal and Vertical distance since those two variables were highly correlated
- Soil types 7, 15, 8 and 25 removed
- Hillshade at 3PM removed

$$X' = \frac{X - \mu}{\sigma}$$

Non categorical variables Normalized

# Modeling

Baseline - Predict tree type based on the most common tree per soil type

Baseline accuracy = 34.94%

# Classifiers

| Classifier | Best Parameters | Dev Set Accuracy | Test Set Accuracy |
|---|---|---|---|
| KNN | Neighbors = 1, Weights = uniform | 82.01% | 68.80% |
| Decision Trees | Criterion = gini, max depth = 17, min_sample_split =2 | 78.17% | 65.65% |
| SVM | C=100, kernel =rbf | 78.44% | 37.05% |
| Logistic Regression | C=10, Penalty = l2 | 66.47% | 3.24% |

# Classifiers

| Classifier | Best Parameters | Dev Set Accuracy | Test Set Accuracy |
|---|---|---|---|
| Extra Trees | Criterion = entropy,  min_sample_split =3, n_estimators = 250 | 87.43% | 77.02% |
| Decision Trees + Adaboost | n_estimators=250 | 86.17% | 76.35% |
| Random Forest | Criterion = gini, n_estimators=500 | 85.58% | 75.16% |
| XGboost | n_estimators=808,learning_rate=0.23, max_depth=10 etc. | 86.11% | 74.46% |

# Lesson 2:

Read the documentation

# Lesson 3:

Think of machine learning as a process

# And our winning model : StackingClassifier

Base Models

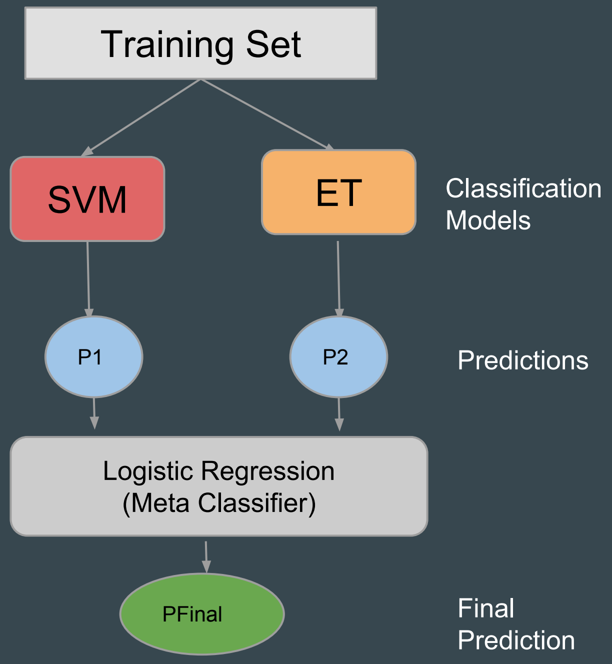
- Extra Trees
- SVM

Stacking Method

- StackingClassifier with best SVM and best ET
- Best LR as meta classifier

Dev set accuracy-  0.878

Test set accuracy- 77.349%



# Lesson 4:

Try radically different techniques

# Lessons Learned

1. Don’t underestimate data visualization
2. Read the documentation
3. Think of machine learning as a process
4. Try radically different techniques
5. Finally prepare for long wait times