STAT 441 Presentation

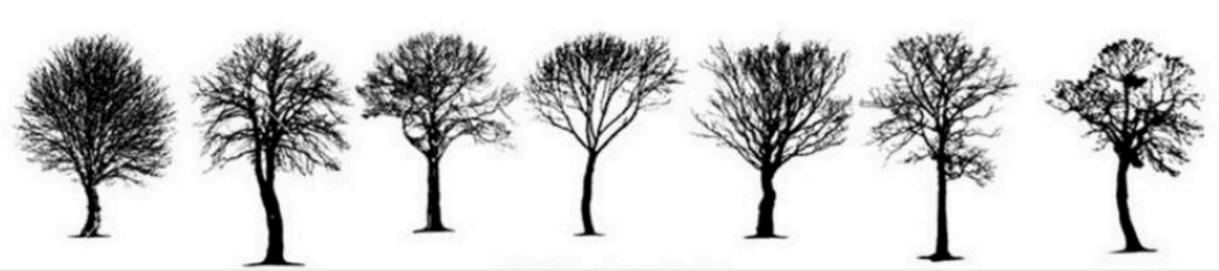
Past Kaggle Competition

Forest Cover Type Prediction

Group 18

Yilun (Tom) Zhang

Yanbing (Miranda) Jin



Outline

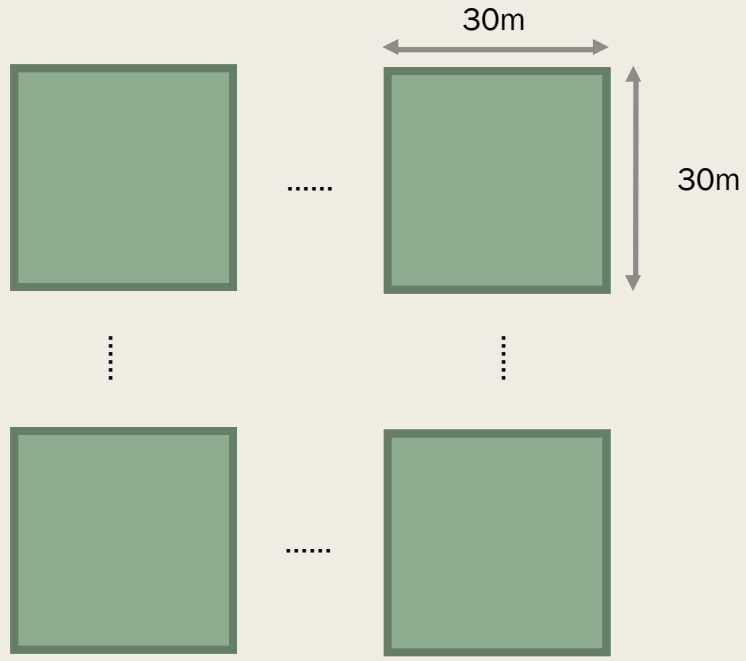
- Data and Problem
- Exploratory Data Analysis
- PCA and LDA
- Supervised Learning Methods
  - Model Fitting
  - Parameter Optimization
  - Feature Creation & Feature Selection
- Semi-supervised Learning Method

Data and Problem

- **Location:** Roosevelt National Forest of northern Colorado
- **Data Source:** US Geological Survey (USGS) and US Forest Service (USFS)

- **Training Set:** 15,120 observations
- **Testing Set:** 565,892 observations
- **Observation:** 30m X 30m patch

- **Goal:** predict the predominant kind of tree cover

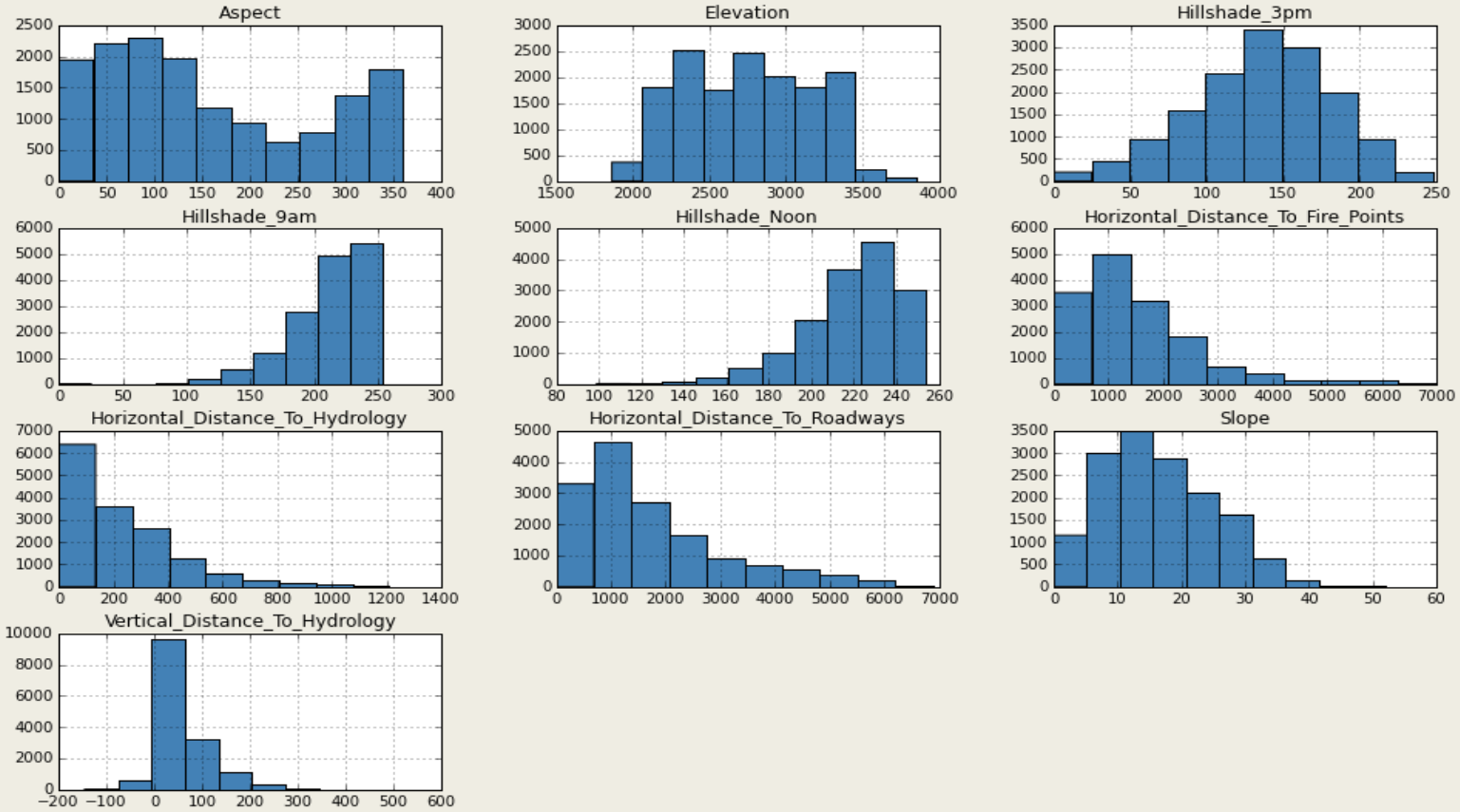


Data and Problem

- **Predictors (12):**
  - **Numerical:** elevation, aspect, slope, horizontal/vertical distance to hydrology, horizontal distance to road/wildfire ignition points, hillshade at 9AM-12PM-3PM
  - **Indicator:** wilderness area designation (4) and soil type (40)
- **Labels (7):**
  - *Spruce/Fir*
  - *Lodgepole Pine*
  - *Ponderose Pine*
  - *Cottonwood/Willow*
  - *Aspen*
  - *Douglas-fir*
  - *Krummholz*

Exploratory Data Analysis – Distribution of Predictors

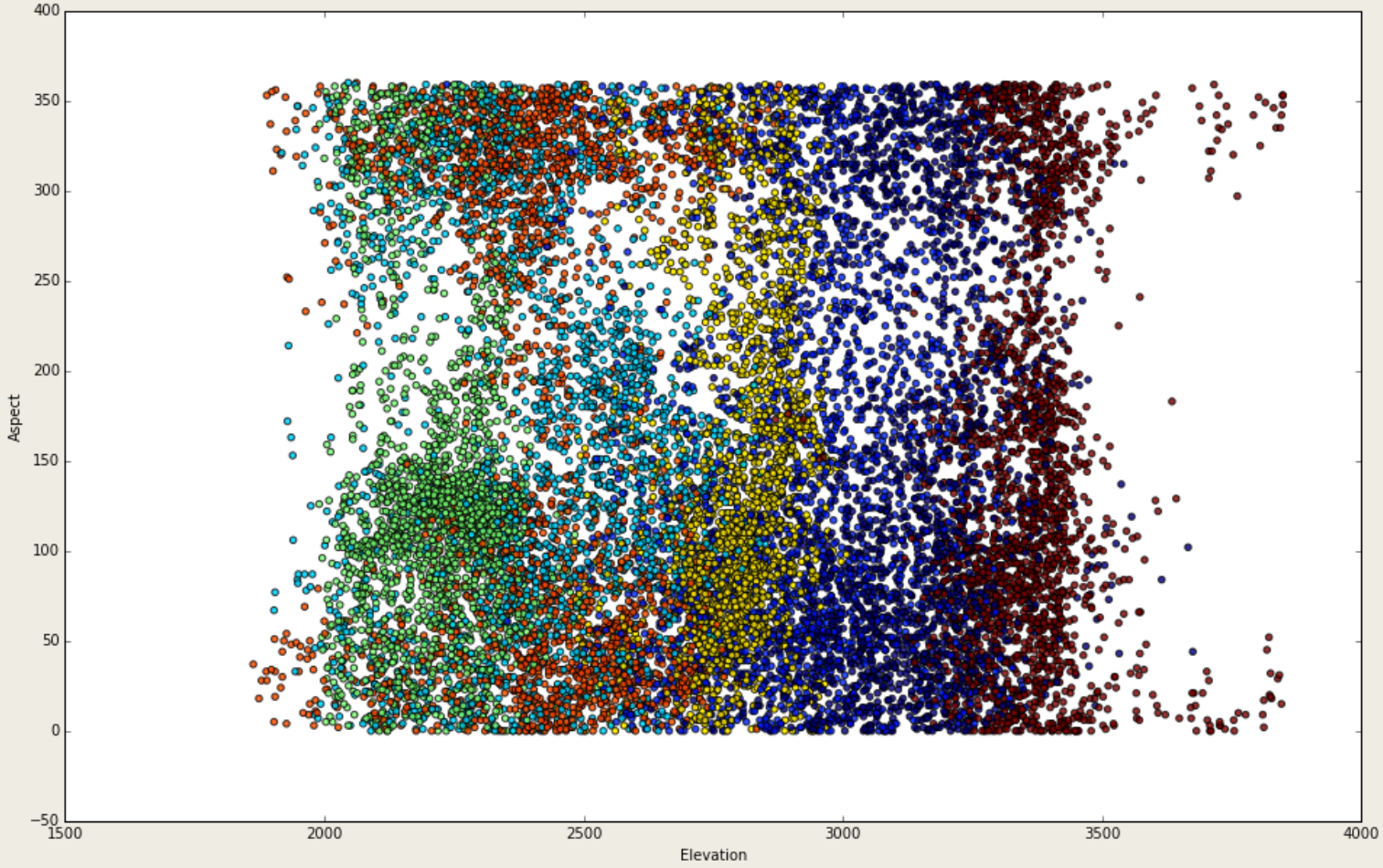
- Histograms of all the numerical predictors



- We might want to **normalize** and **transform** some of the variables so that they look normal

Exploratory Data Analysis – Scatter Plots Between Predictors

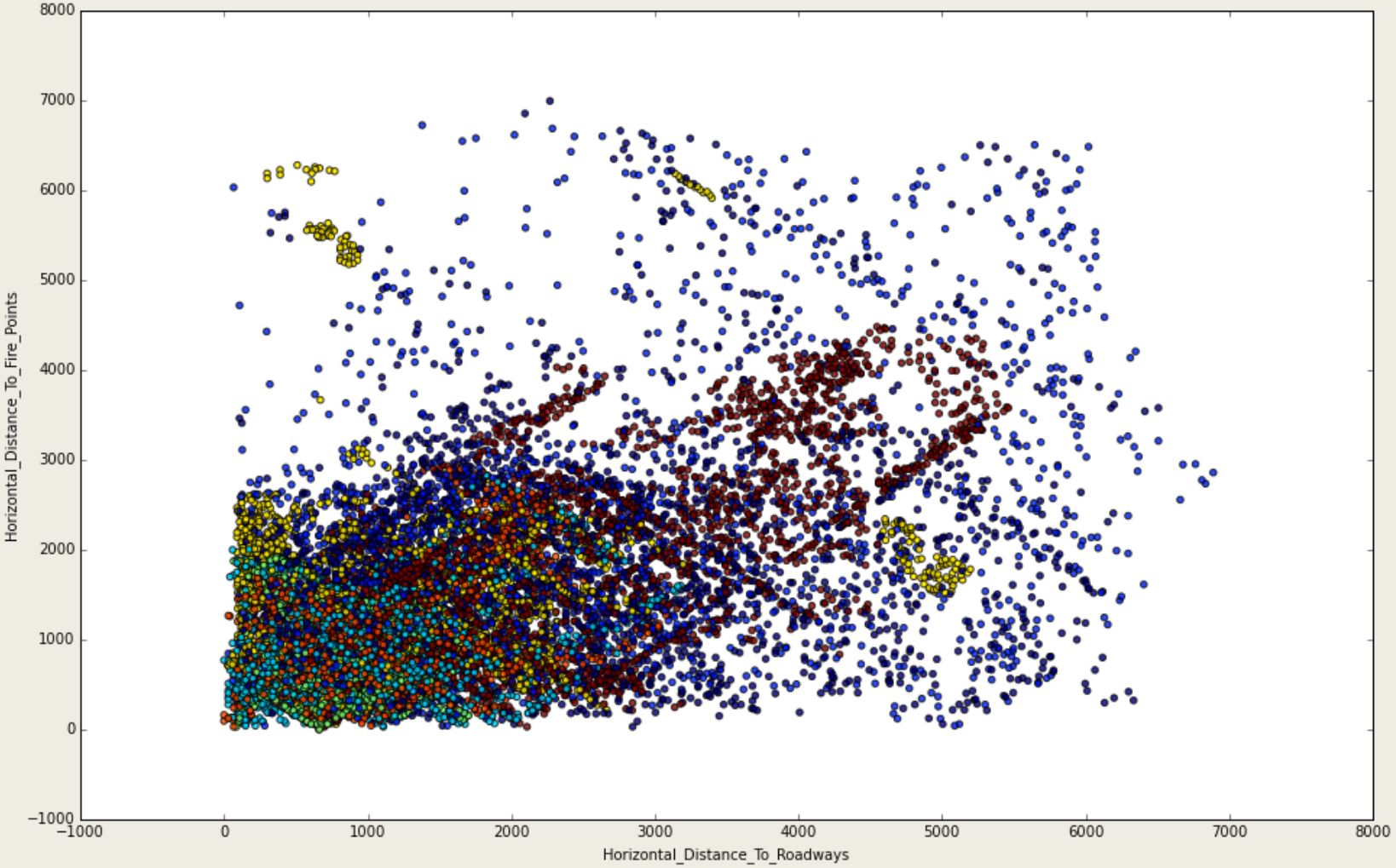
- Selected scatterplots of Aspect vs. Elevation, colored by labels



- Predictors can be **significant** when predicting the labels

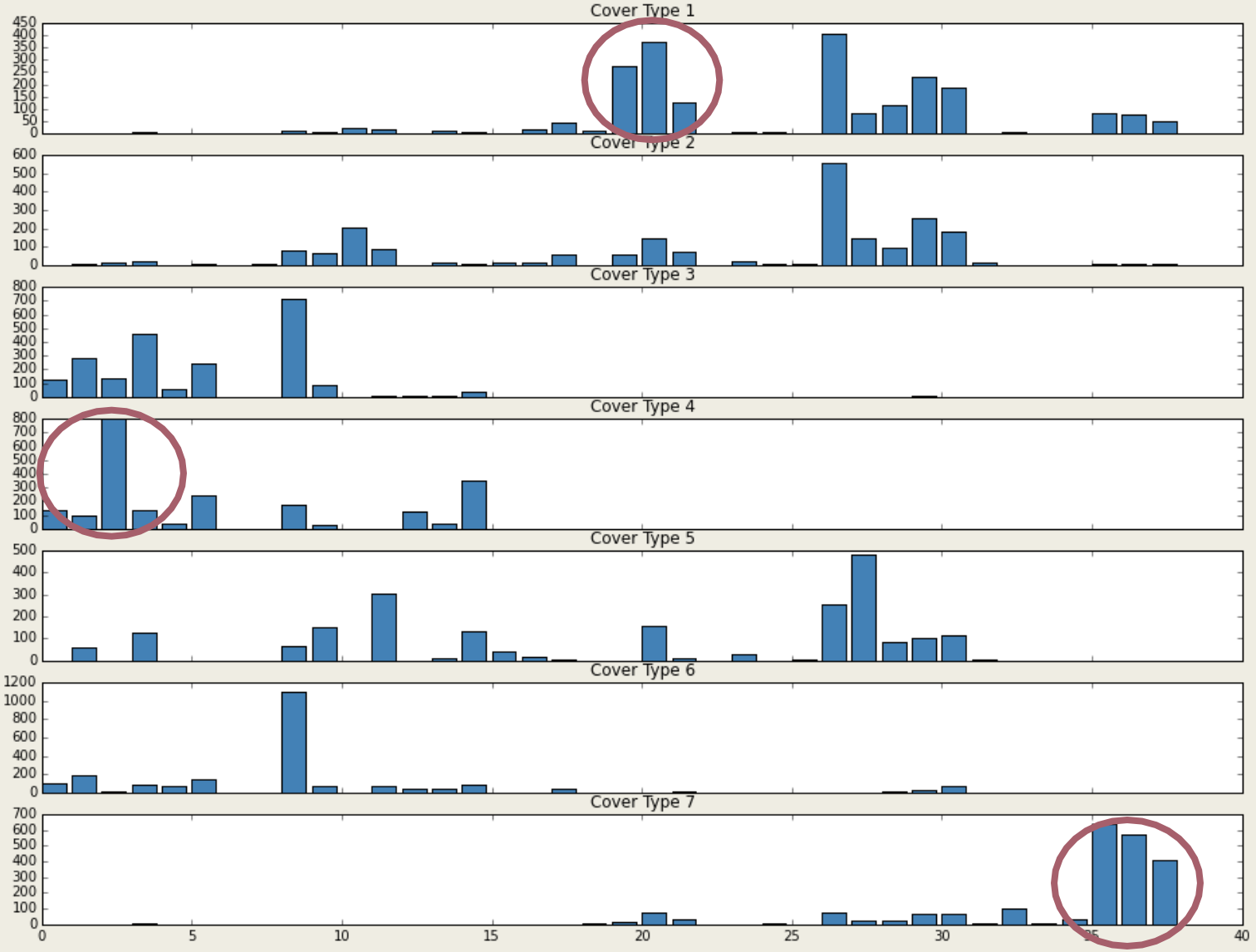
Exploratory Data Analysis – Scatter Plots Between Predictors

- Selected scatterplots of H_Dist_Fire_Points vs. H_Dist_Roadways, colored by labels



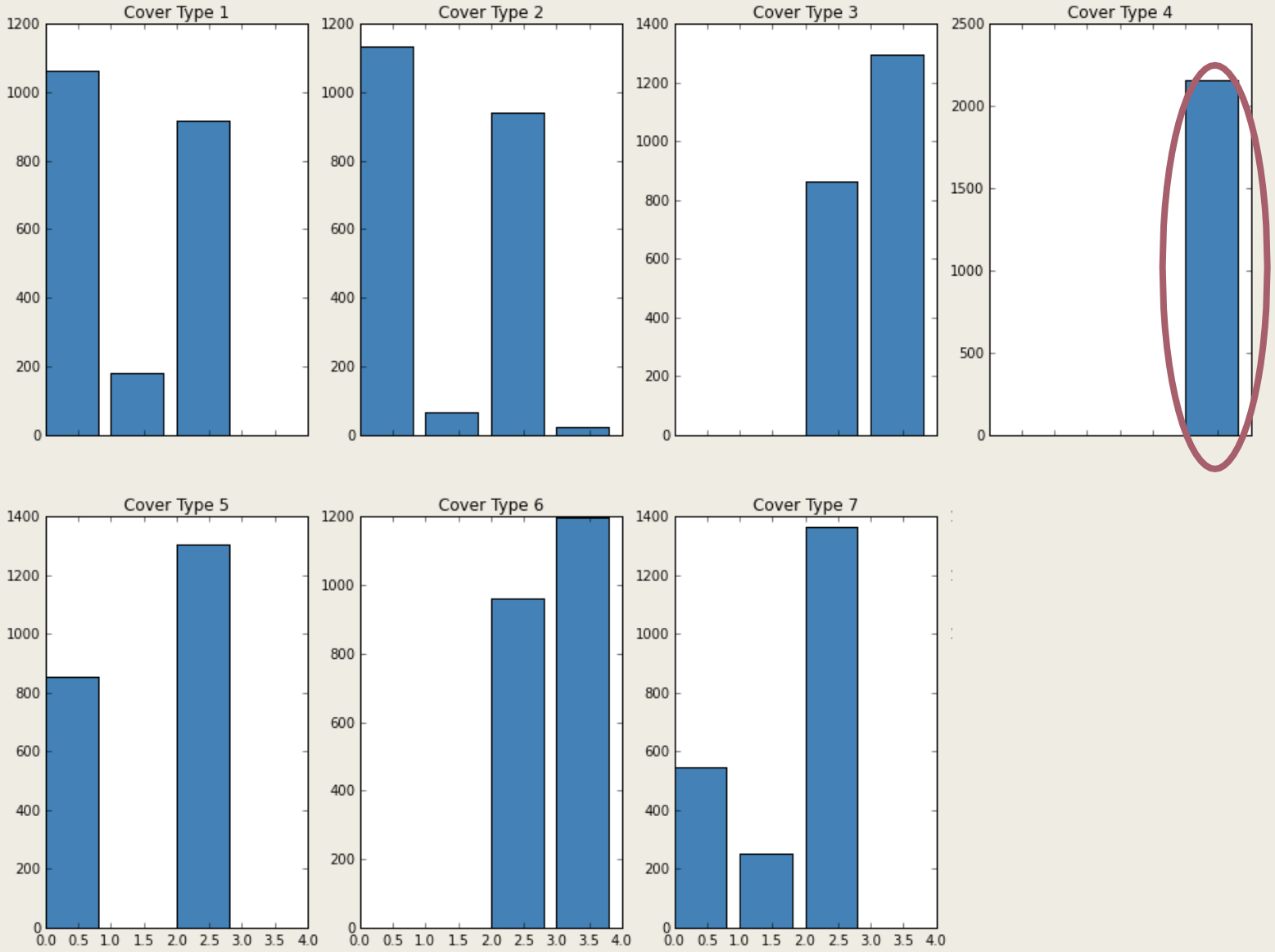
Exploratory Data Analysis – Relationship with Labels

- Soil Types



Exploratory Data Analysis – Relationship with Labels

- Wilderness Types



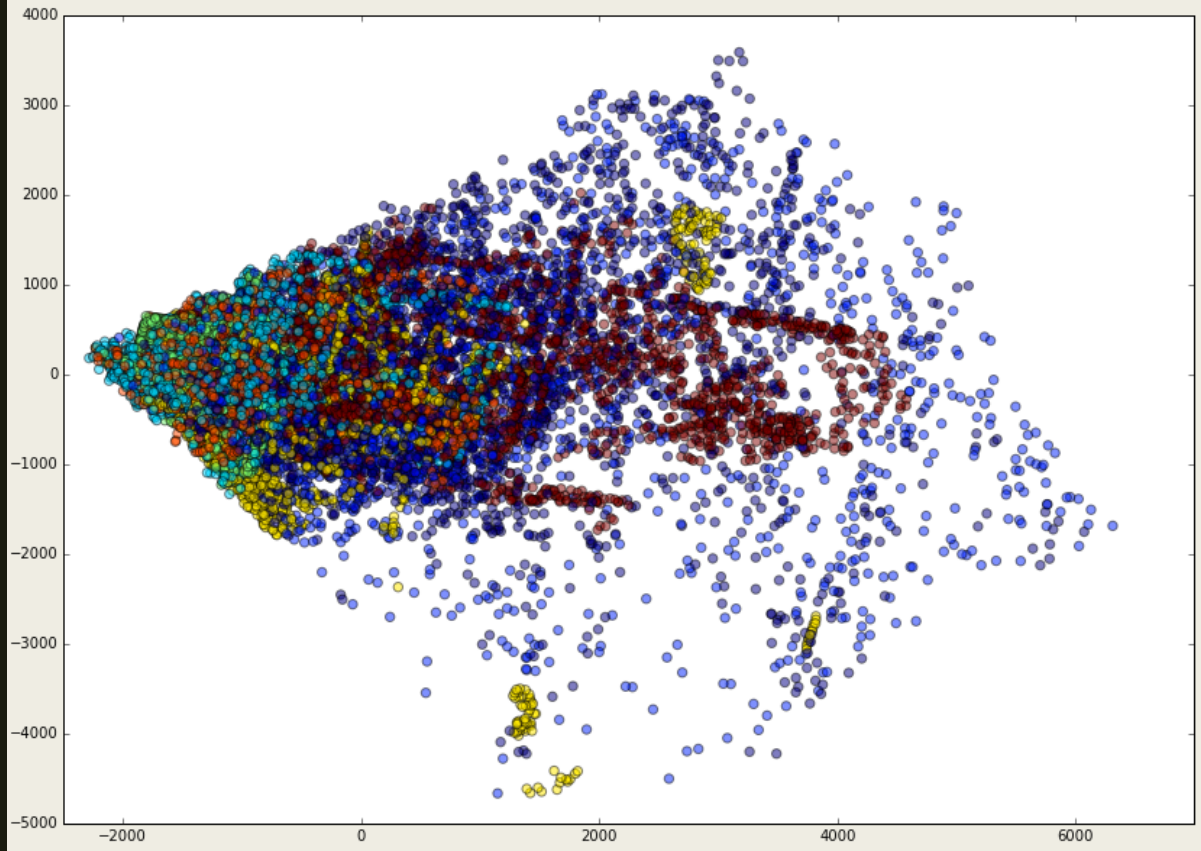
Exploratory Data Analysis

- Labels (Cover Type)

Label	Count
1	2160
2	2160
3	2160
4	2160
5	2160
6	2160
7	2160

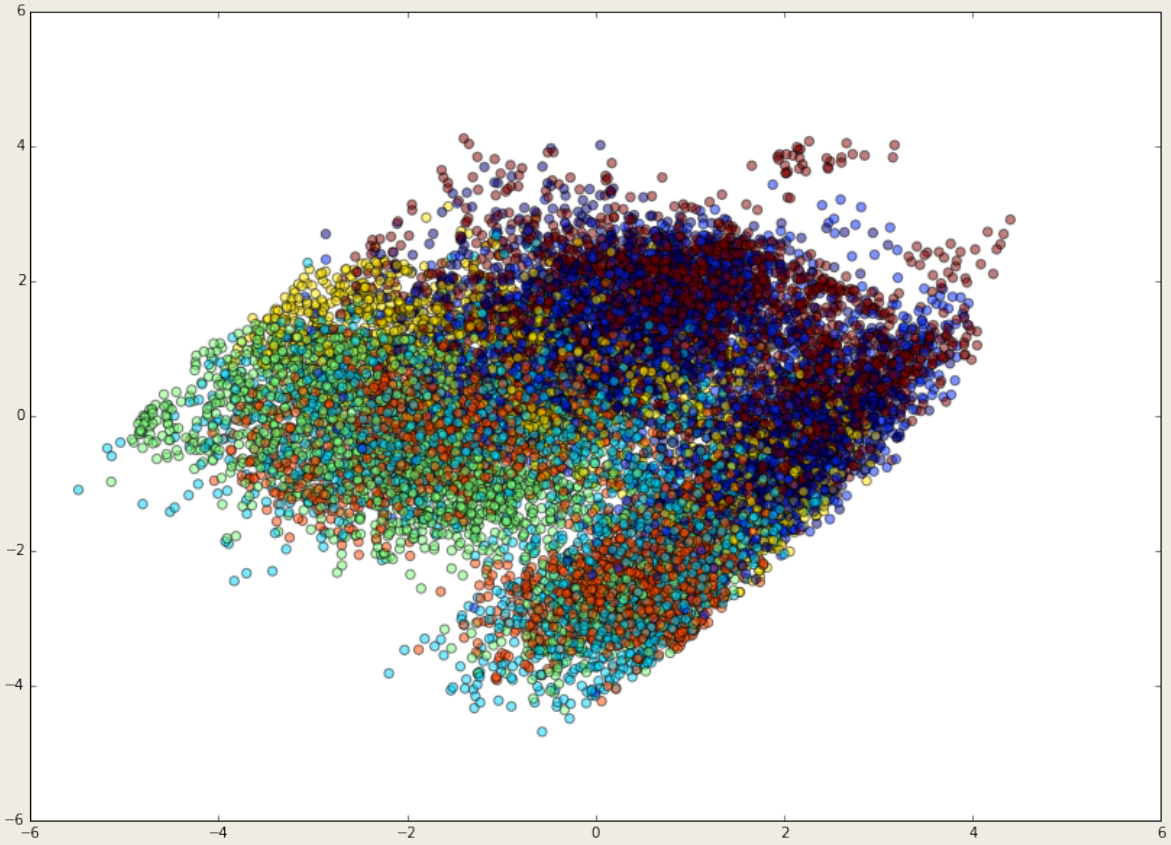
- Balanced classes

PCA – 2D



Without Transformation

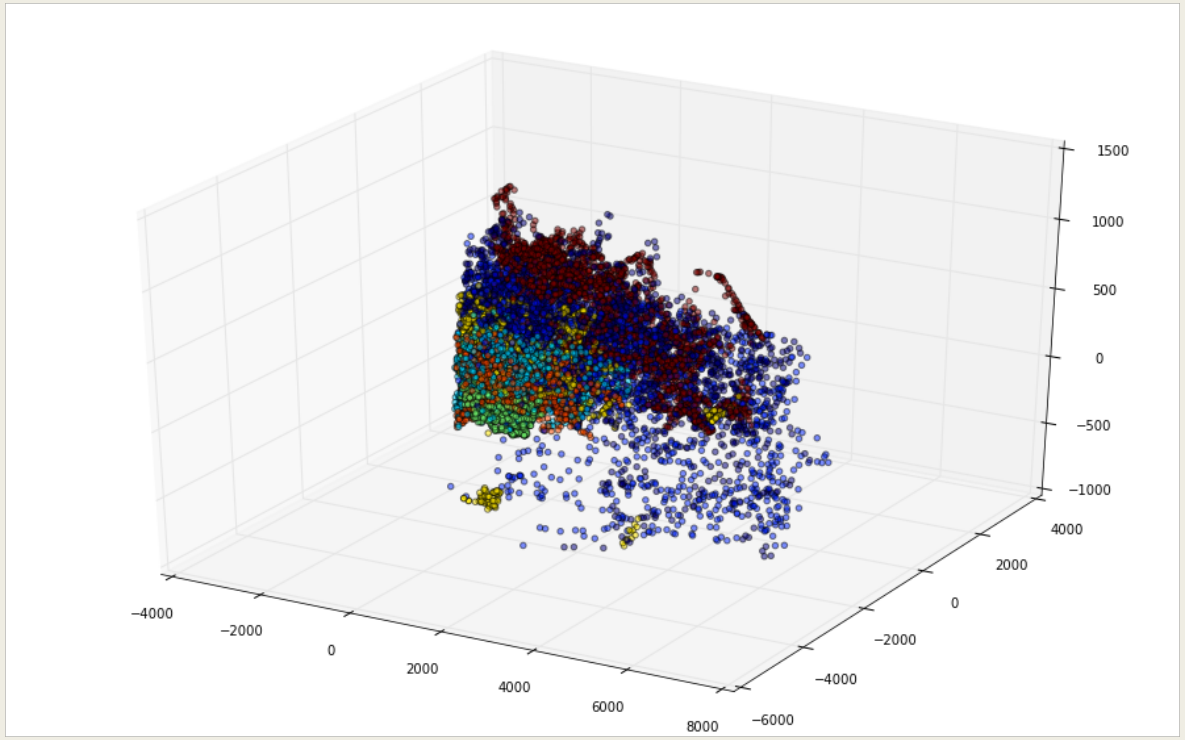
94.76%



With Transformation

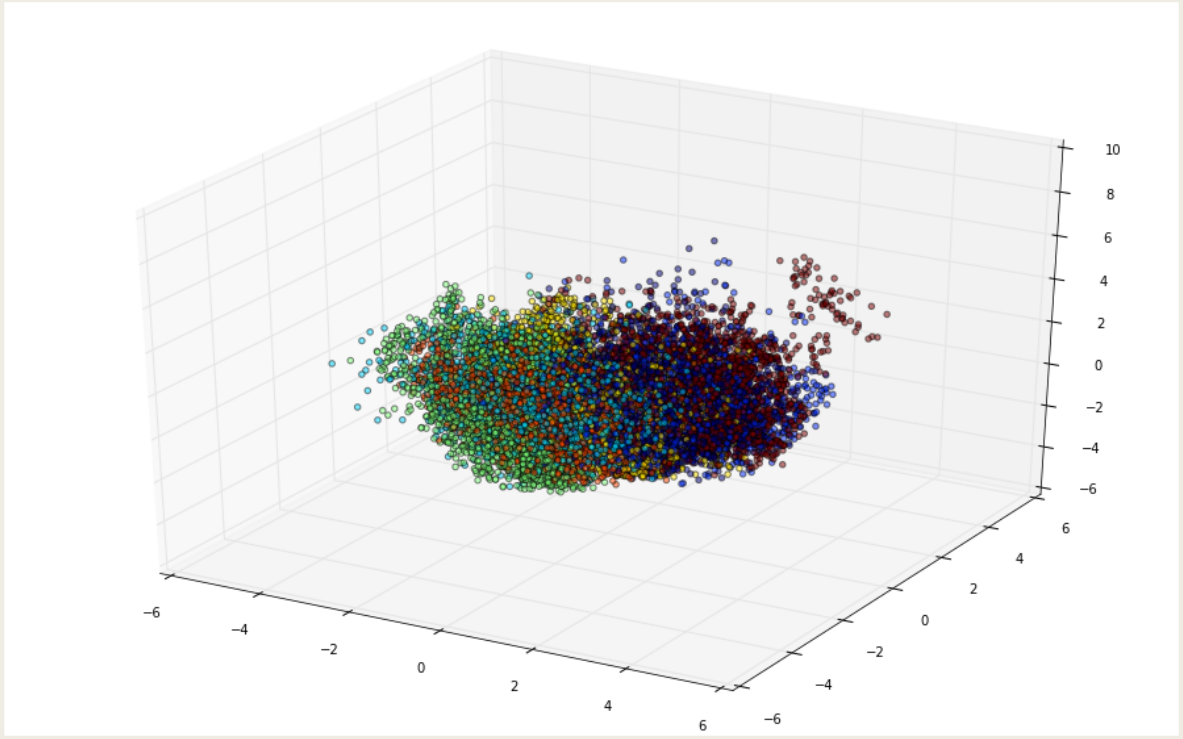
45.76%

PCA – 3D



Without Transformation

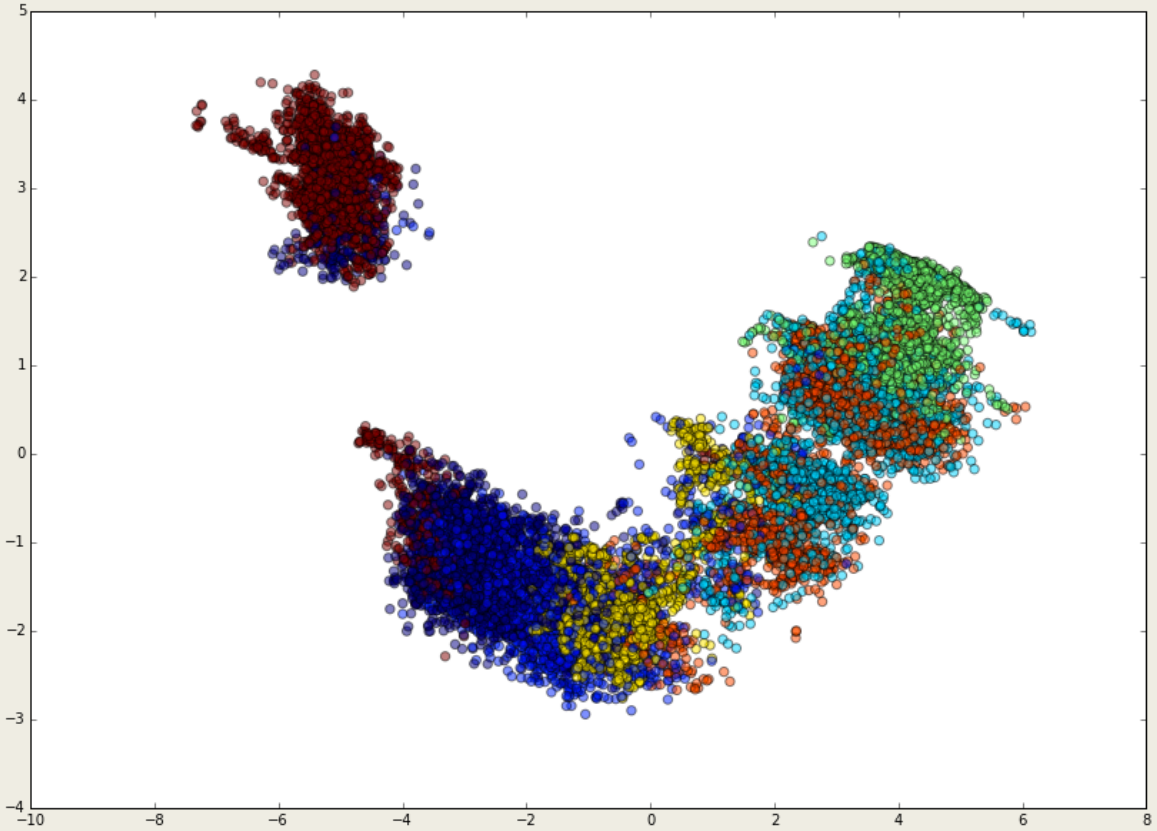
98.35%

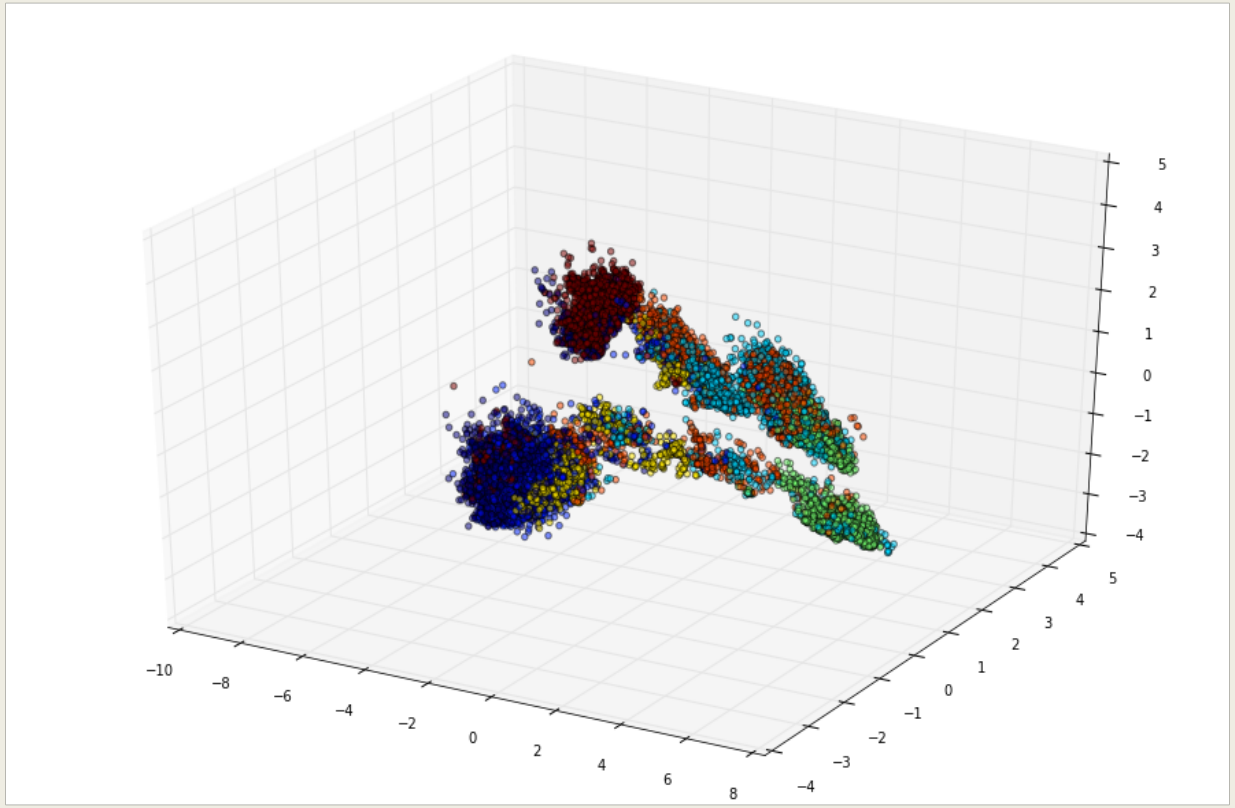


With Transformation

61.27%

LDA – 2D & 3D


5
4
3
2
1
0
−1
−2
−3
−4
−10
−8
−6
−4
−2
0
2
4
6
8


5
4
3
2
1
0
−1
−2
−3
−4
−10
−8
−6
−4
−2
0
2
4
6
8
5
4
3
2
1
0
−1
−2
−3
−4

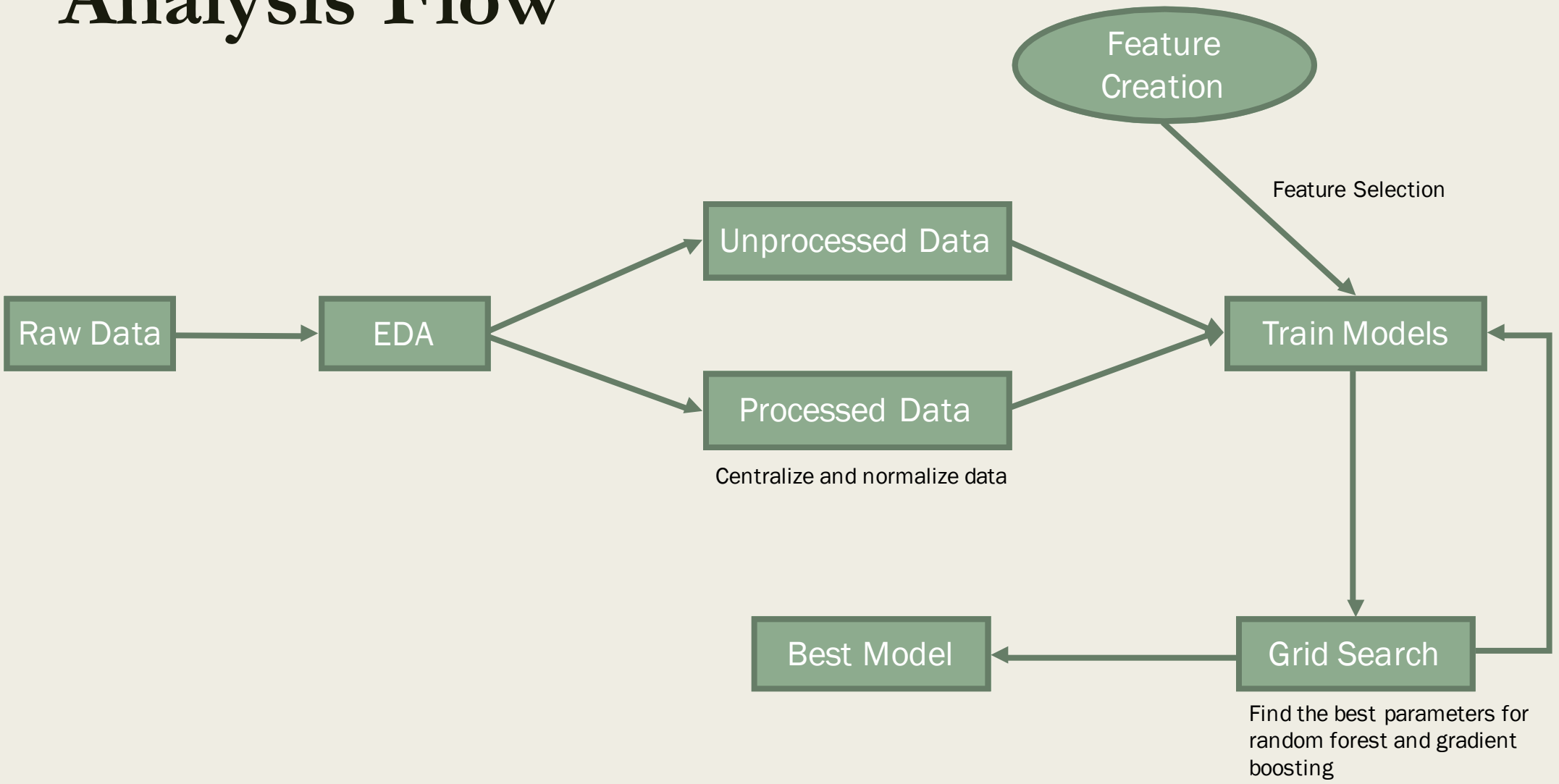
Supervised Learning Methods

- Multiclass Logistic Regression
- Tree Methods
  - Decision Tree
  - Random Forest
  - Extremely Randomized Trees
- Gradient Boosting (Trees)
- SVM
- Gaussian Naïve Bayes
- Adaptive Boosting
- Neural Network

Pros and Cons

Model Name	Pros	Cons
Logistic Regression	Fast, less prone to over-fitting Easy interpretation	When data is not linearly separable
Decision Tree	Fast	Might overfit the training data
Random Forest Extremely Randomized Trees	Robust to over-fitting, fast, takes categorical values	Interpretation
Gradient Boosting	Might overfit	Interpretation, much slower than RF
SVM	Can model complex non-linear relationship Robust to noise	Time, memory, tuning, interpretation
Gaussian Naïve Bayes	Less model complexity Requires less data	Assumes class conditional independence
Adaptive Boosting	Iteratively update weights on weak classifiers	Sensitive to noise data
Neural Network	Works good on data with non-linear relationship	Prone to over-fitting, time, interpretation Complexity, training data size

Analysis Flow



Model Error Rate with Default Parameters

Model	On Raw Data	On Transformed Data
Logistic Regression (LR)	0.3227	0.3122
Decision Tree (DT)	0.1991	0.1984
Random Forest (RF)	0.1634	0.1614
Extremely Randomized Trees (ERT)	0.1468 	0.1521
Gradient Boosting (GB)	0.1944 	0.1948
SVM	0.5390	0.3168
Gaussian Naïve Bayes (GNB)	0.5324	0.5304
Adaptive Boosting (AB)	0.5728	0.5727
Neural Network (NN)	0.4008	0.3214

Grid Search - Parameters

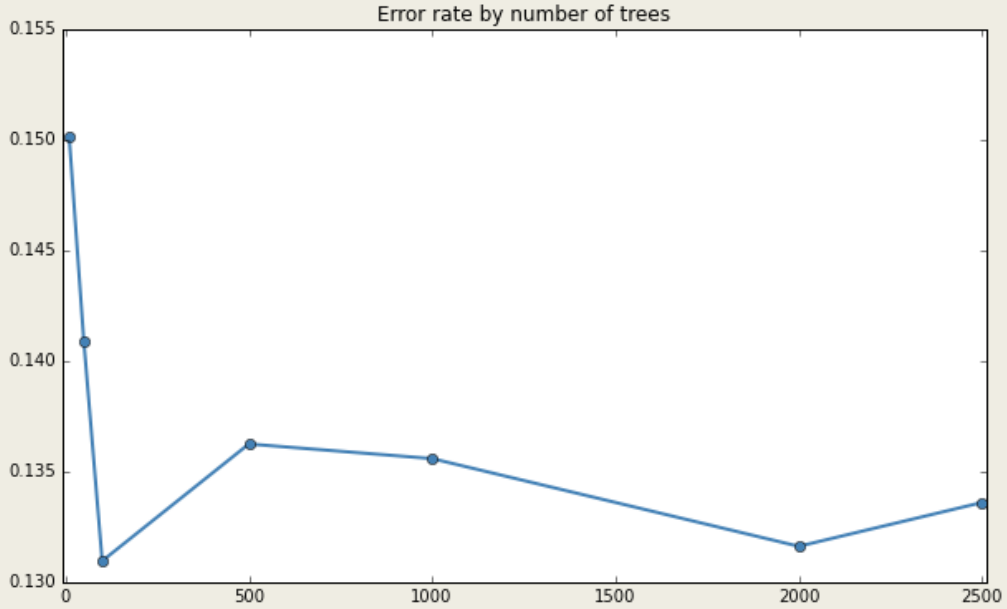
- **Extremely Randomized Trees**

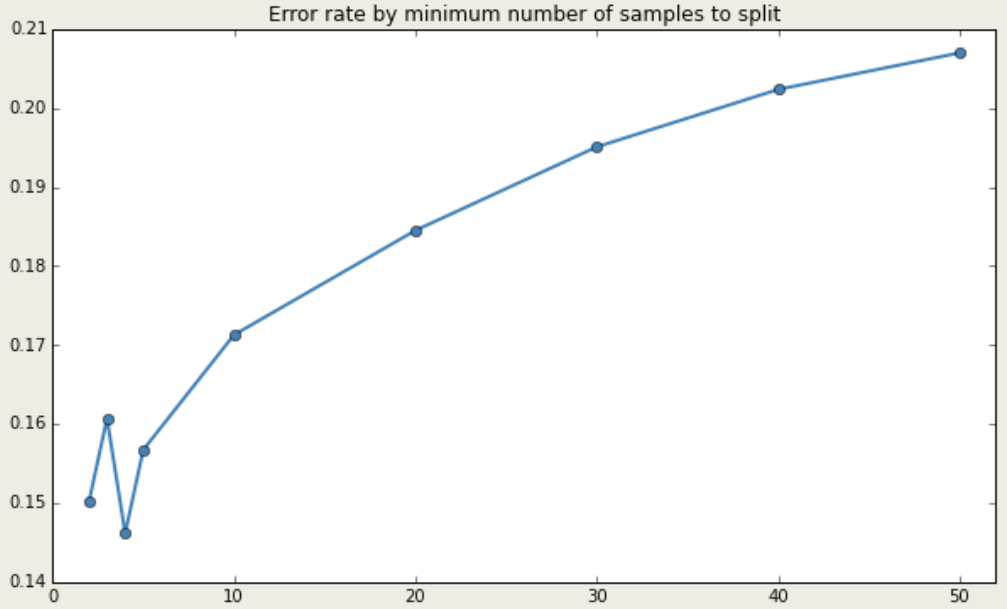
Parameter	Values
Number of estimators	[10,50,100,500,1000,2000,2500]
Minimum sample split	[2,3,4,5,10,20,30,40,50]
Split quality measure criterion	["gini","entropy"]
Bootstrap sampling	[False, True]
Warm start (add more estimators to previous trees or generate a new one)	[False, True]

- **Gradient Boosting**

Parameter	Values
Number of estimators	[10,50,100,200,500,1000]
Minimum sample split	[2,3,4,5,10,20,30,40,50]
Learning rate	[0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1]

Grid Search – ERT





Parameter	False	True
Bootstrap	0.1501	0.1726

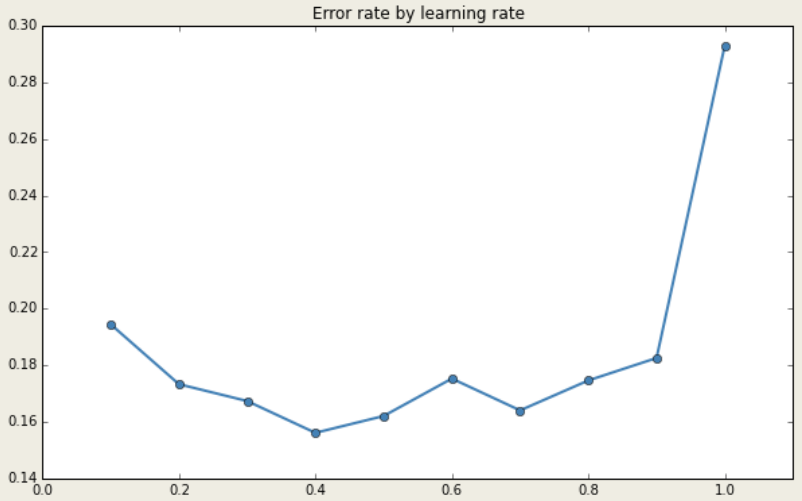
Parameter	gini	entropy
Criterion	0.1501	0.1448

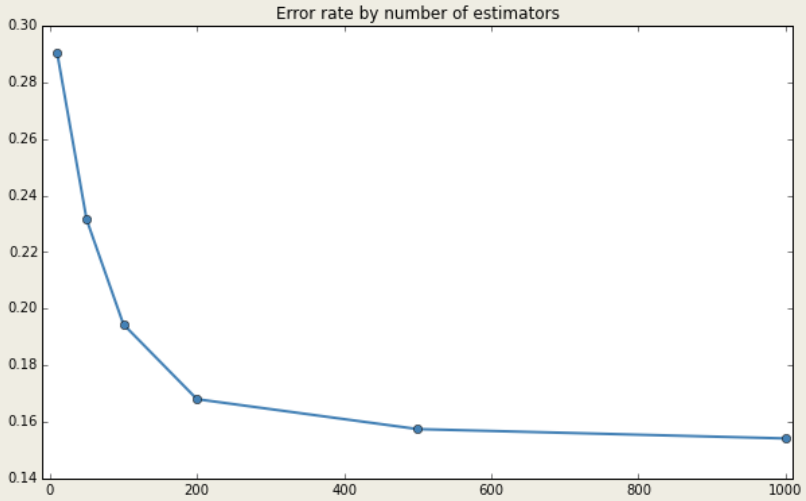
Parameter	False	True
Warm Start	0.1501	0.1501

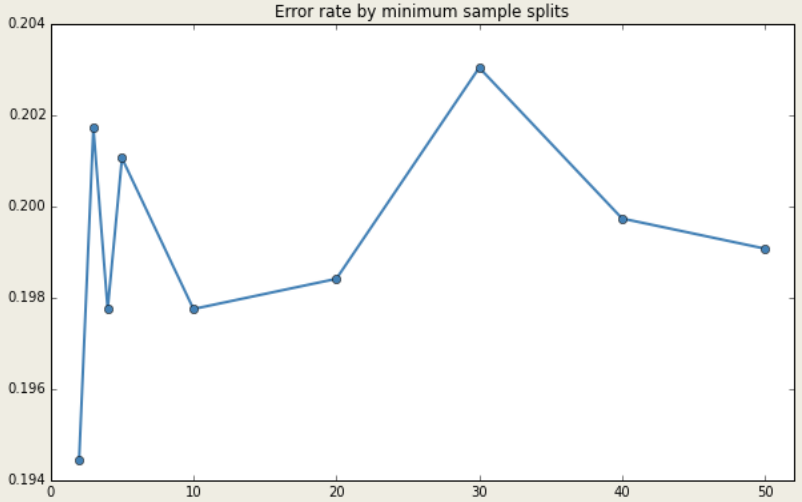
Number of Estimator	Min Sample Split	Bootstrap	Criterion	Error
100	2	False	gini	0.1309

0.0159

Grid Search – GB







Number of Estimator	Min Sample Split	Learning Rate	Error
2000	3	0.3	0.1402

0.0542

Parameter Creation

- **OPTION 1**
- 2-way interaction between all 10 numerical predictors (excluding **Wilderness_AreaX** and **Soil_TypeX**)
- Elevation^2, Elevation*Aspect, Elevation*Slope, …, Hillshade_Noon*Hillshade_3pm
- We now have **109** predictors

- **OPTION 2**
- 3-way interaction between all 10 numerical predictors
- Elevation^3, Elevation^2*Aspect, Elevation^2*Slope, …, Hillshade_3pm^3
- We now have **329** predictors

Grid Search – ERT Interaction

- 2-way Interaction

Number of Estimator	Min Sample Split	Bootstrap	Criterion	Error
2000	2	False	gini	0.1124

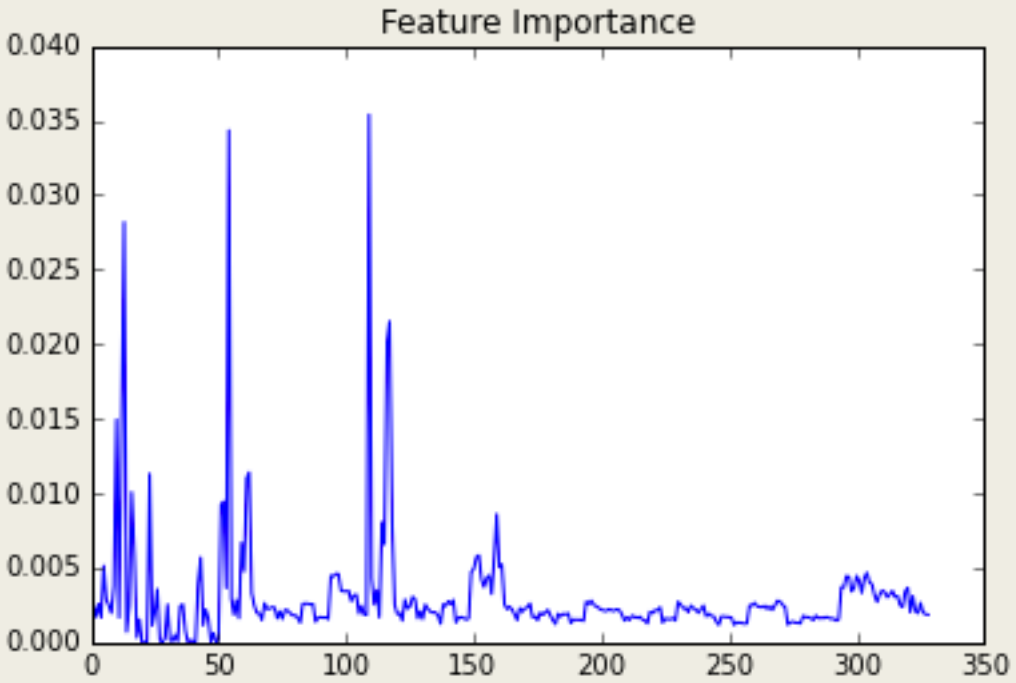
0.0185

- 3-way Interaction

Number of Estimator	Min Sample Split	Bootstrap	Criterion	Error
100	4	False	gini	0.1230

0.0079

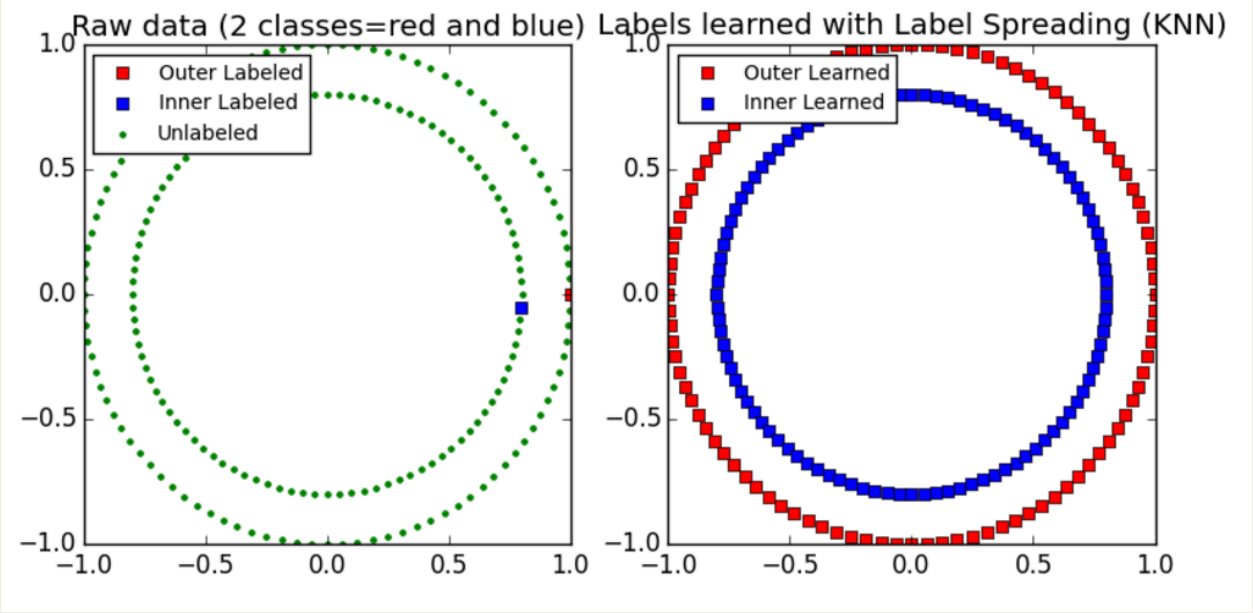
3-Way Interaction – Feature Importance

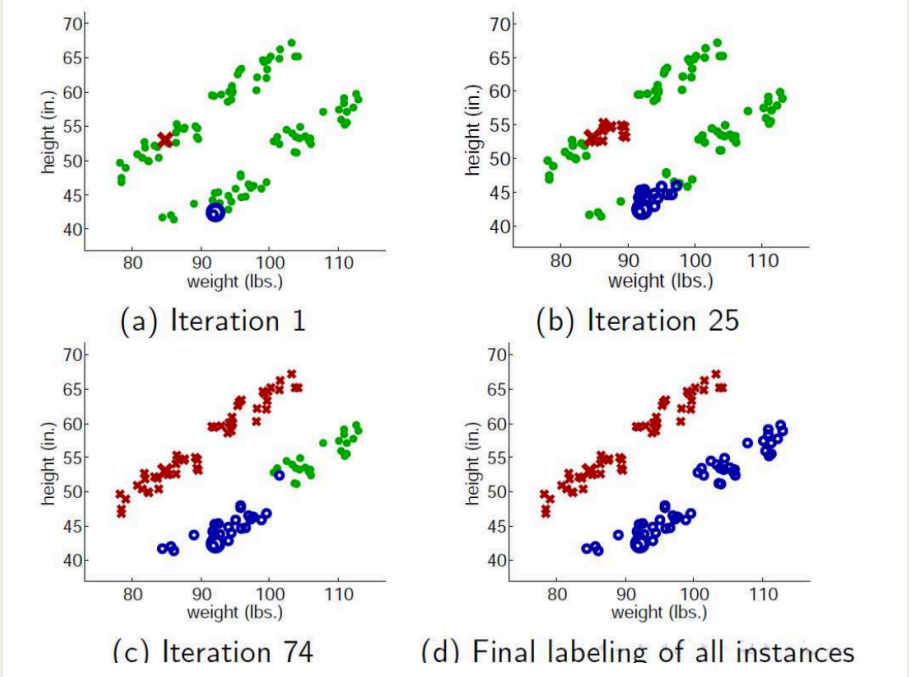


- Select variables with importance > 0.0025
- Total of 116 features
- Test error: 0.1197 ↓ 0.0033

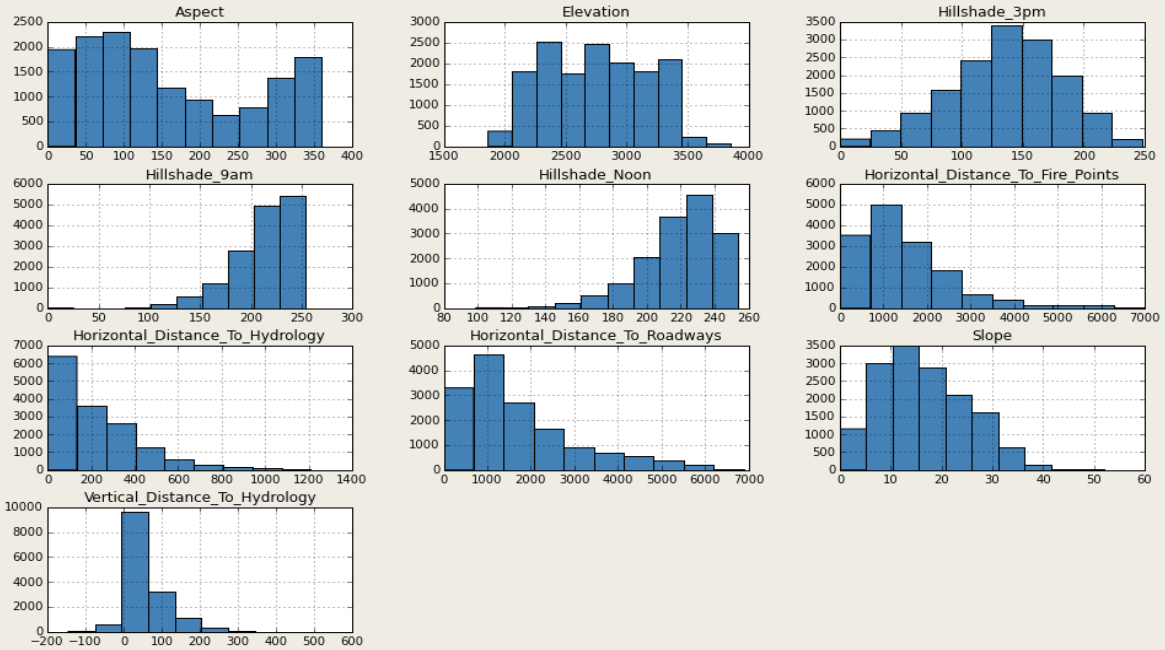
Semi-Supervised Learning Methods

- **Idea**
  - Training set size 15,120; Testing set size 565,892
  - Borrow data from testing set
  - Better capture the shape of the underlying data distribution and generalize better to new samples

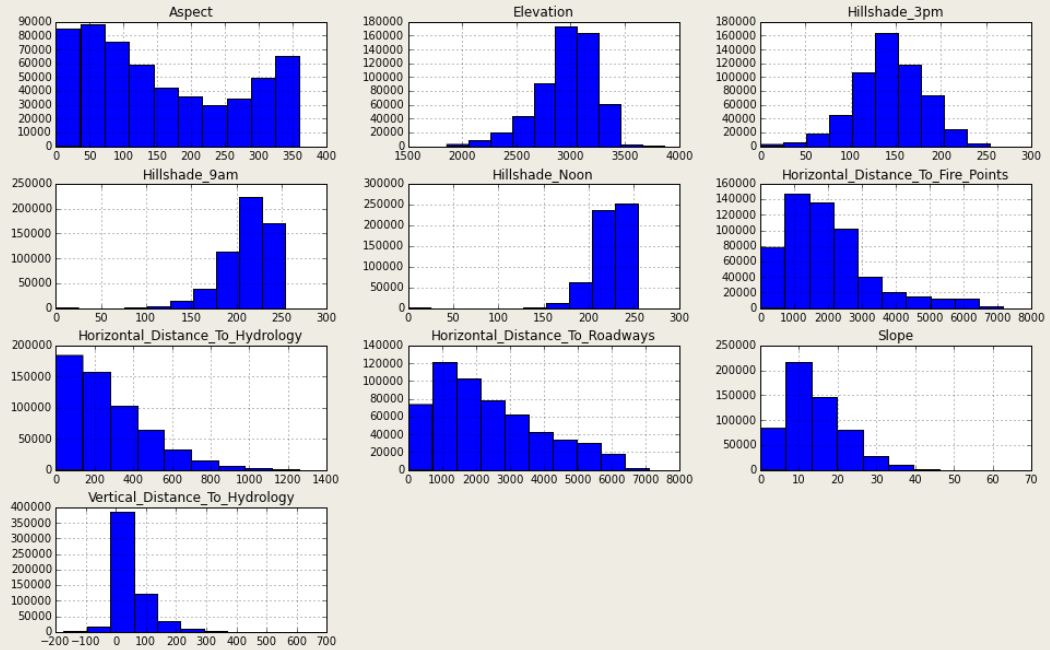




Distribution of Numerical Variables



Training Data



Testing Data

Semi-Supervised Learning Methods

- **Steps:**
  - Borrow X% data from testing set randomly (set a random state)
  - Label Propagation / Label Spreading
  - Use the original training data + predicted labels as training set to train model
    - **ERT** for now
  - Predict on testing data

- Also:
  - Change borrow rate (5%, 10%, 20%, etc.)
  - Fit other possible models
  - Use original data vs. data with interaction columns

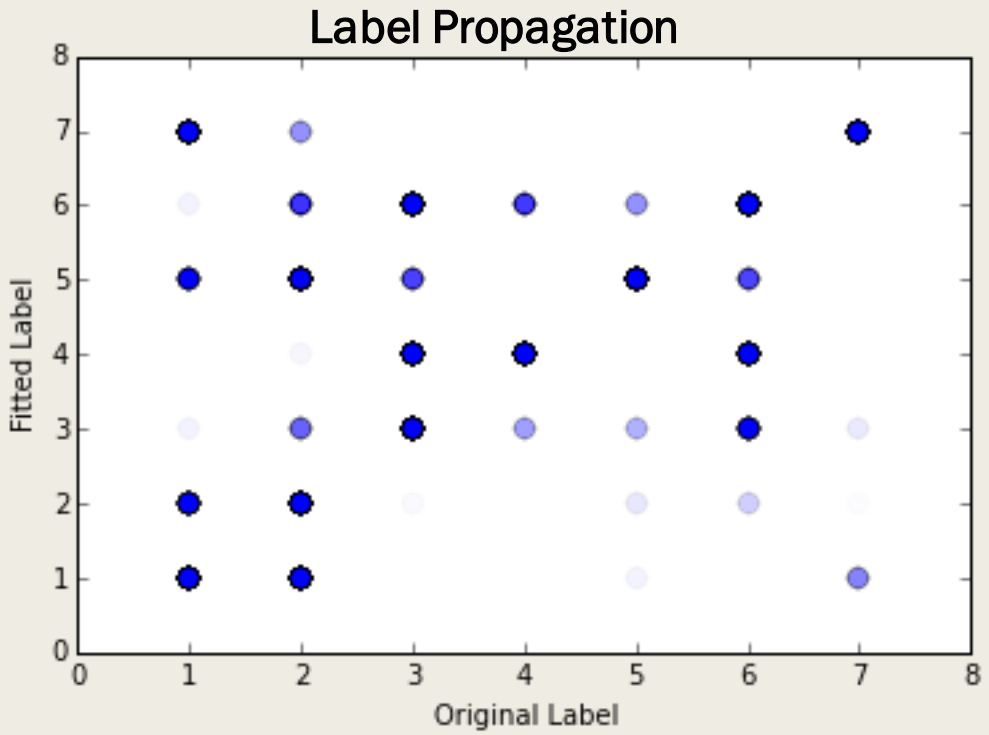
Semi-Supervised Learning Methods

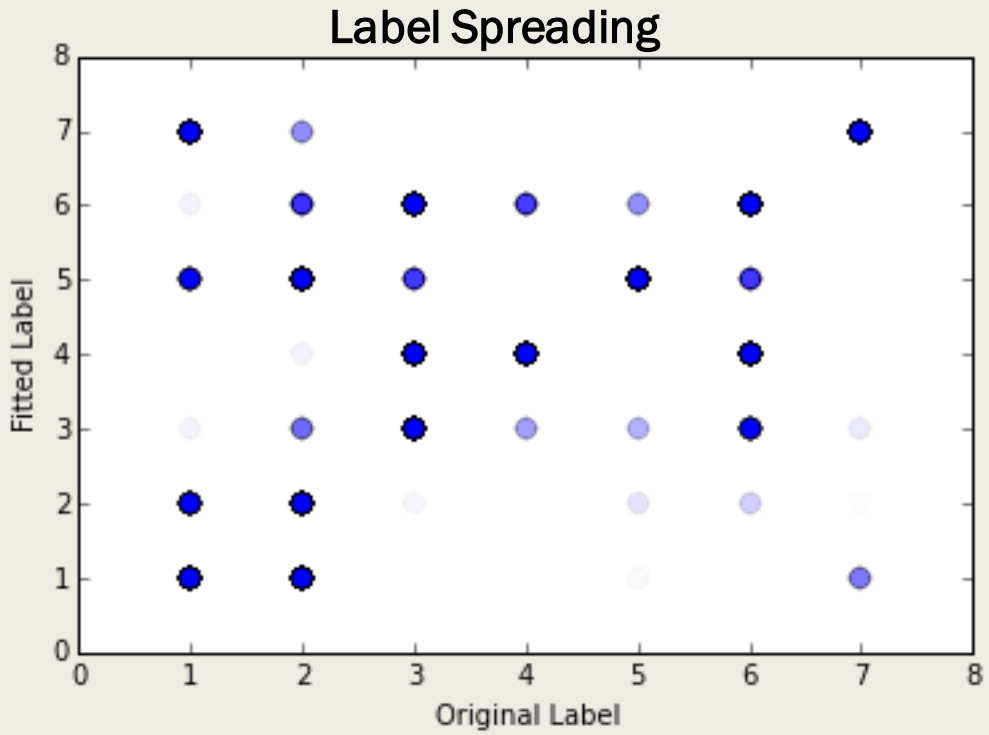
- the similarity matrix to clamp the label distributions is different
- Clamping allows the algorithm to change the weight of the true ground labeled data distribution to some degree ($\boldsymbol{\alpha}$ level)
  - Hard clamping ($\boldsymbol{\alpha} = 1$), can be modified
- Kernel: { RBF, KNN }

- Label Propagation:
  - Uses raw similarity matrix constructed from data
- Label Spreading:
  - Minimizes a loss function that has regularization properties (robust to noise)
  - On each iteration, it modifies the matrix and normalize the edge weights by calculating the normalized graph Laplacian matrix

Semi-Supervised Learning Methods

- The label propagation / spreading algorithm changes the original label of the training data (~ 20%)
- Change them back





Validation Result – Raw Data

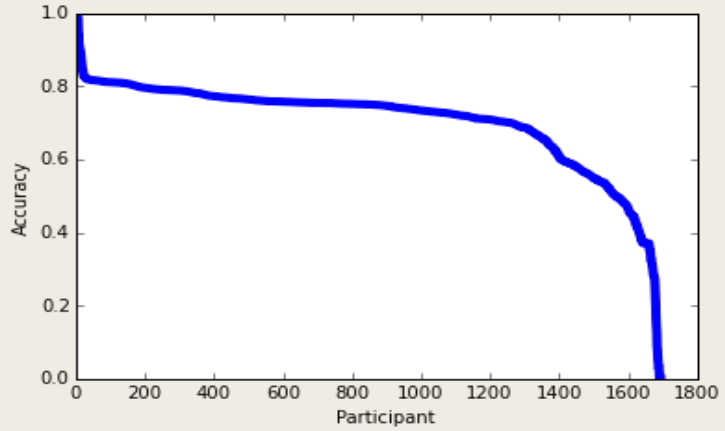
- Label Spreading method
- KNN kernel only
- Default ERT model with 1000 estimator

% Test Data Borrowed	Alpha	Testing Error
5%	1.0	0.1865
5%	0.8	0.1865
5%	0.6	0.1889
10%	1.0	0.1386
10%	0.8	0.1534
10%	0.6	0.1590
20%	1.0	0.0959
20%	0.8	0.1062
20%	0.6	0.1101

Label Propagation method

% Test Data Borrowed	Alpha	Testing Error
5%	1.0	0.1852
5%	0.8	0.1909
5%	0.6	0.2015
10%	1.0	0.1403
10%	0.8	0.1527
10%	0.6	0.1692
20%	1.0	0.0969
20%	0.8	0.1058
20%	0.6	0.1183

Kaggle Result



Model	Score	Rank (out of 1694)
ERT + raw data	0.75793	591
GB + raw data	0.72622	1074
ERT + 2-way interaction	0.77803	368
ERT + 3-way interaction	0.75972	540
ERT + 3-way interaction + best feature importance filter	0.77985	362
Semi-Sup Label Spreading + raw data + borrow 20% + alpha 1.0 + ERT	0.65952	1349
Semi-Sup Label Spreading + raw data + borrow 50% + alpha 1.0 + ERT	0.73131	1022
Semi-Sup Label Spreading + 2-way interaction data + borrow 50% + alpha 1.0 + ERT	0.73125	1022
Semi-Sup Label Spreading + 2-way interaction data + borrow 100% + alpha 1.0 + ERT	0.77320	395

Thank you