Forest Cover Type Prediction

Final Project W207.5 Summer 2015

Nate Black, Vineet Gangwar, Jared Maslin, and Malini Mittal

Competition Summary

- Predicting the cover type label of 30 x 30 meter plots of forest, as defined by the U.S. Forest Service (USFS)
- Independent variables defined from the U.S. Geological Survey, as well as the USFS
- Data hosted by the UCI machine learning repository, with a training set and test set provided
- Study area included four wilderness areas located in the Roosevelt National Forest of northern Colorado
- The goal is to predict an integer classification for the forest cover type

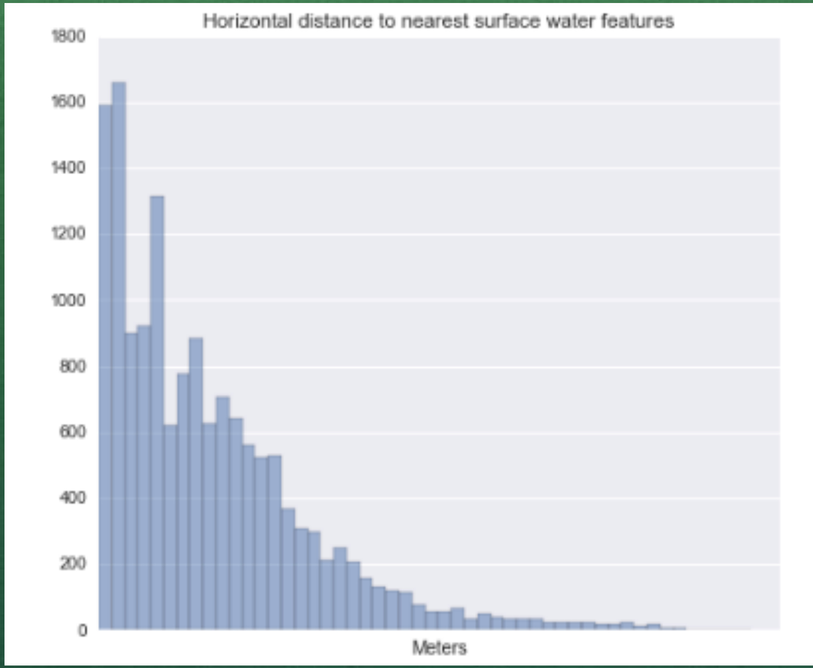
The Dataset

- Primary data fields provided:
  - Elevation (in meters)
  - Aspect (in degrees azimuth)
  - Slope (in degrees)
  - Horizontal distance to nearest surface water features
  - Vertical distance to nearest surface water features
  - Horizontal distance to nearest wildfire ignition points
  - Horizontal distance to nearest roadway
  - Hillshade (0 to 255 index, summer solstice) – At 9am, Noon, and 3pm
  - Wilderness_Area (4 types, 0 = absence or 1 = presence)
  - Soil_Type (40 types, 0 = absence or 1 = presence)
  - Cover_Type (7 types, integers 1 to 7)

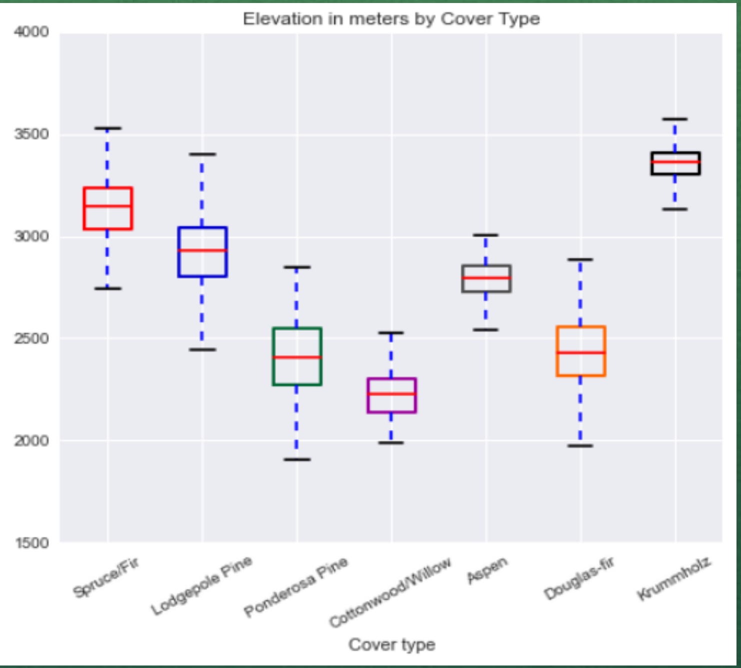
Initial Analysis

- The training data held an even split across all 7 cover types, whereas the test data did not.
  - Test data was heavily skewed toward types 1 and 2
- Boxplots showed a clear divergence in elevation
- Histograms of “hillshade” measures displayed a difference in the behavior of “3pm” measures, in comparison to the “9am” and “noon”
- The 40 soil types could actually be condensed into 11 types

Visualizations Examples



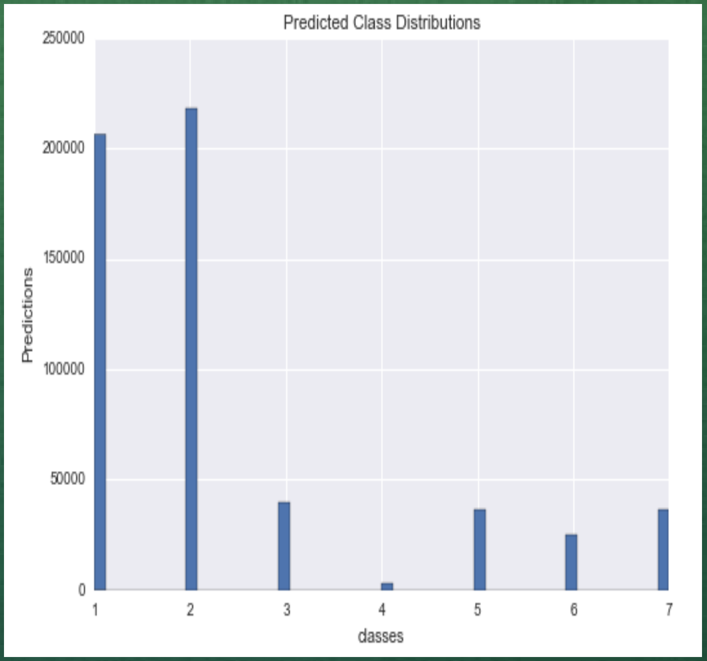
Dispersion of horizontal distance to water across training data observations



Dispersion of elevation across training data cover types

Baseline Results for Kaggle

- A simple KNN (K=1) model, fitted with training data and applied it to the test data, produced a Kaggle score of 0.71016 (rank of 1,175)
- Histogram of predicted values show a heavy presence of cover types 1 and 2 (roughly 75%)



Additional Baseline Validation

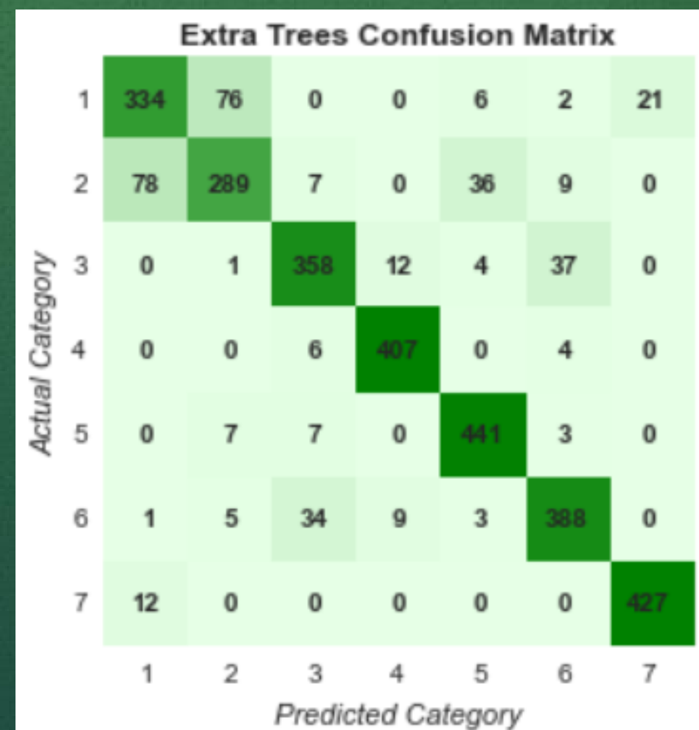
- Development data used for validation; not test
- Additional baseline models were run using an 80%/20% random split from the training set to create a training and development subset
  - KNN (k=1): Accuracy = 79.79%
  - Linear Regression: Accuracy = 41.85%
  - Logistic Regression: Accuracy = 67.00%
  - Gaussian Naïve Bayes: Accuracy = 47.55%
  - Random Forest: Accuracy = 83.04%
  - **Extra Trees: Accuracy = 84.03%**
    - **Winner!!!**

Error Analysis: Extra Trees

- Confusion matrix and classification report both reflect a significant error rate in the prediction of cover types 1 & 2
  - Recurrence of the “skew” between spread of cover types between the training and test datasets
  - More specifically, the two types appear to be mistaken for one another

Classification report:

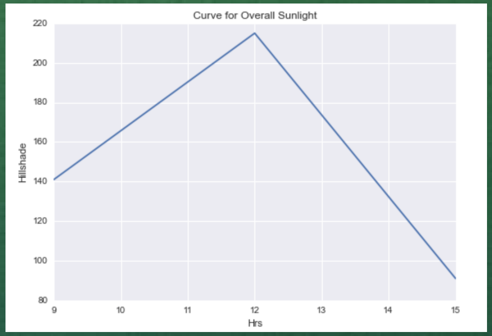
	precision	recall	f1-score	support
1	0.79	0.76	0.77	439
2	0.76	0.69	0.73	419
3	0.87	0.87	0.87	412
4	0.95	0.98	0.96	417
5	0.90	0.96	0.93	458
6	0.88	0.88	0.88	440
7	0.95	0.97	0.96	439
avg / total	0.87	0.87	0.87	3024



Feature Engineering

- Collapsed 40 soil types into 11 types *
- Used class weights to account of skew in test data

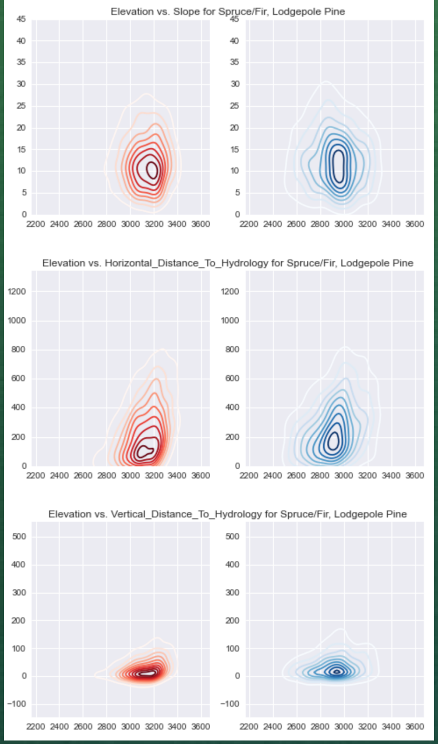
- Added 16 new Engineered Features:
    - Total Energy (AUC) based on hillshade
    - Is Fire point closer than water?
    - Is Roadway (human activity) closer?
    - Euclidean distance to water
    - Water elevation higher than tree's

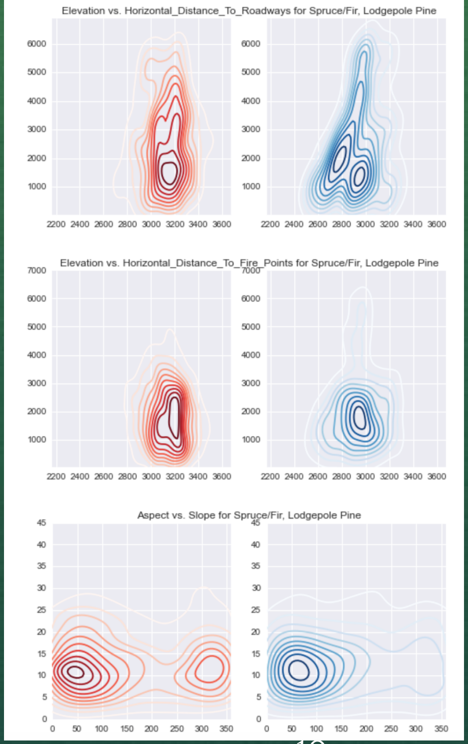


* https://archive.ics.uci.edu/ml/machine-learning-databases/covtype/covtype.info

Feature Engineering – Amplifying variation

In addition to the test being skewed towards Cover_Types 1 and 2, these two cover types are also very similar:





To amplify variation we created 2 new features by multiplying features:

- Slope X Horizontal distance to water X Vertical distance to water
- Distance to roadways X Distance to fire sources (product of distance to man-made features)

Results and Conclusion

- ExtraTreesClassifier with RandomizedSearchCV to ascertain best parameters
- Improved our baseline accuracy from 0.71016 to 0.80403 with a Kaggle rank of 163
- Multiplying features were improving the accuracy but was probably overfitting
- Domain knowledge about trees would help
- Skewed nature of test data could be explored more