FOREST COVER TYPE:

CLASSIFYING FORESTS FROM CARTOGRAPHIC VARIABLES

MELIA MILLER

INTRODUCTION: PROBLEM STATEMENT

- Determining forest cover types is expensive or not feasible for large or remote forest areas
    - direct observation
    - remote sensing
- By using machine learning algorithms, we can predict forest cover type from cartographic variables
- By predicting forest cover types, we can better understand what factors influence cover types in a given area

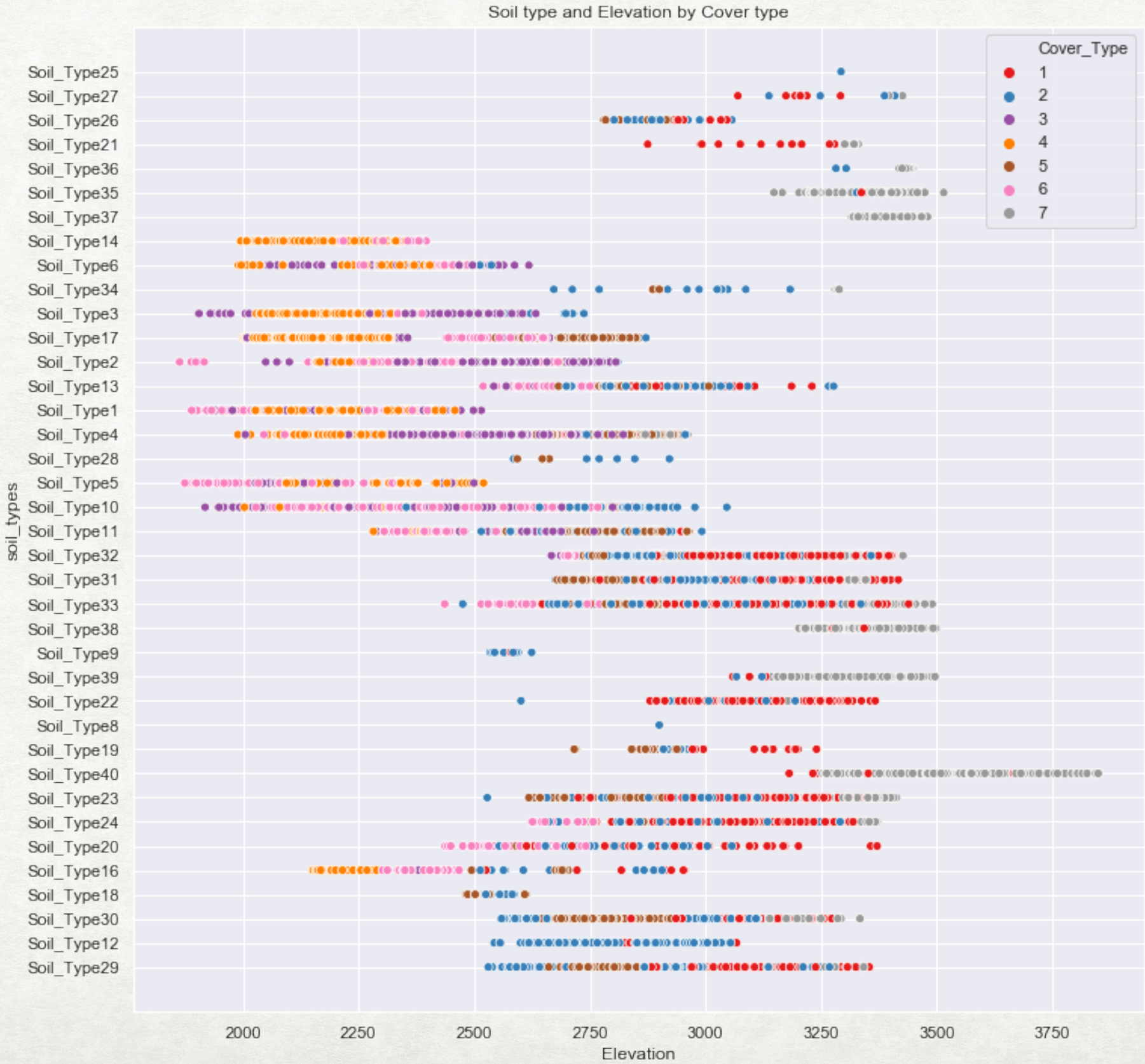
INTRODUCTION: VALUABLE SOLUTION

- Machine learning solutions can save organizations time and money spent on directly observing forest types
- Accurate predictions of forest cover types can be used to track forest health over time and determine inventory levels in logging
    - Global impacts on climate and commerce

INTRODUCTION: DATASET

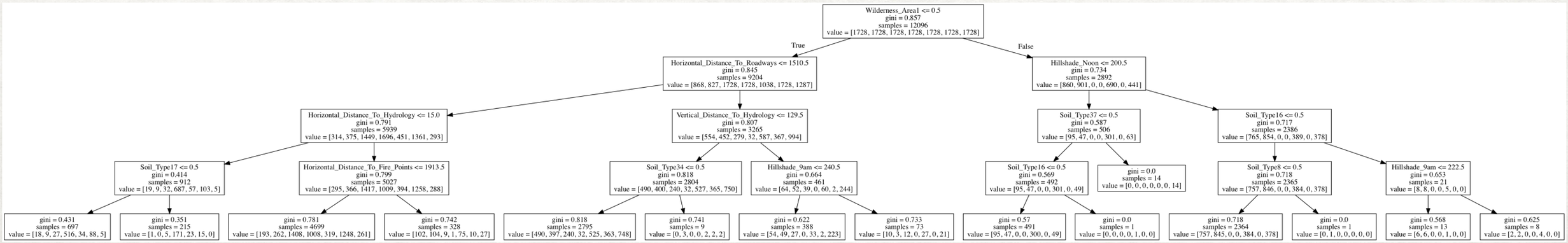
- This data has been aggregated from US Geological Survey and US Forest Service (USFS) Region 2 Resource Information System data
- Forest cover types include Spruce/Fir, Lodgepole Pine, Ponderosa Pine, Cottonwood/Willow, Aspen, Douglas-fir, and Krummholz
- The dataset was put together by UCI and was used as a Kaggle competition 2015
- The data files can be downloaded from Kaggle
- The data set as 15,120 rows with 2,160 samples for each forest cover type
    - Samples are 30 meter by 30 meter cells

STAGE OF DATA SCIENCE PROCESS: EXPLORATION



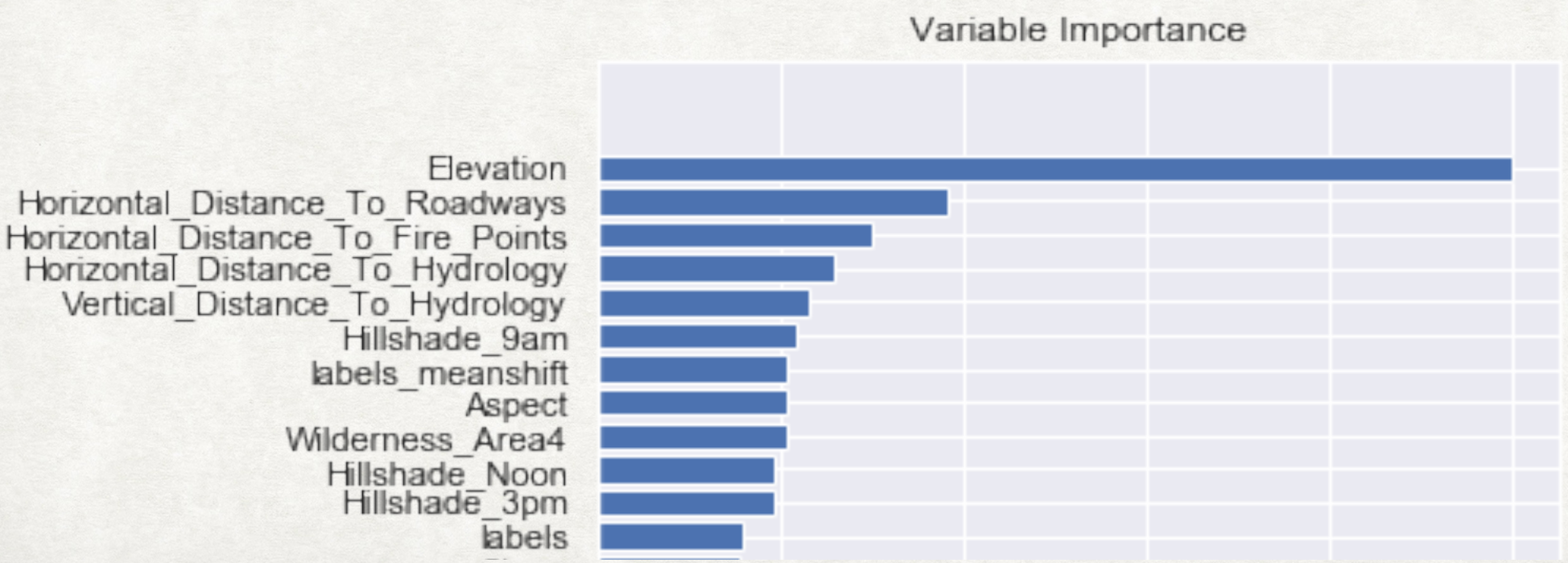
STAGE OF DATA SCIENCE PROCESS: BUILDING BASELINE

- Prepare data for modeling, split into training and testing groups, stratify
- Build models - overall scores
  - k nearest neighbors - 0.80
  - support vector machine - 0.72
  - random forest - 0.86
  - logistic regression - 0.70



STAGE OF DATA SCIENCE PROCESS: FEATURE ENGINEERING

- Unsupervised learning
  - k-means and mean shift to get clusters
  - soil type text create features using natural language processing
  - using these features random forest now score 0.87



STAGE OF DATA SCIENCE PROCESS: ADVANCED MODELS AND COMPARISON

- Keras/tensorflow to build neural network
- Types of neural network:
  - multilayer perceptron
  - convolutional neural network

```
Epoch 100/100
12096/12096 [==============================]
c: 0.8221
--- 358.5493907928467 seconds ---
Test loss: 0.01766992320439645
Test accuracy: 0.8220899470899471
```

- random forest takes about 3 seconds - accuracy score is 0.87

Random forest

```
2    2018
1      75
5      32
6      18
3      13
7       4
Name: y_pred, dtype: int64
```

MOVING INTO PRODUCTION ENVIRONMENT

- Use model to predict forest types in other locations and use other inputs such as images and user inputted variables
- Create database to store inputs and maintain database
- Build models using new data and track model performance over time, update model monthly with new data - more data should improve model accuracy
- Further work to improve predictions based on land management organizations needs - help make informed decisions on resources

FUTURE WORK - - future is now

- Main focus - improve model performance
  - Another pipeline: start with binary predictions for class 2
  - Repeat binary prediction for any class with low f1-score
  - Use predictions as features and add to model — results in 0.97 accuracy scores using MLP neural net - - add train, validation, test

- Create simple interactive user interface where user can upload data and visualize 3D plots of variables

SOURCES

- https://pdfs.semanticscholar.org/42fd/f2999c46babe535974e14375fbb224445757.pdf
- https://www.kaggle.com/c/forest-cover-type-kernels-only
- https://archive.ics.uci.edu/ml/datasets/covertype
- http://cs229.stanford.edu/proj2014/Kevin%20Crain,%20Graham%20Davis,%20Classifying%20Forest%20Cover%20Type%20using%20Cartographic%20Features.pdf
-

THANK YOU!

- Koyuki Nakamori

- QUESTIONS?