# The Problem

- Prediction of Forest Cover Type
- Original number of variables/features- 12
- Training Data Set- 15120 examples, 56 columns (2 features one hot encoded which increased dimensionality)
- Test Data set- 565892 values to predict

# Training Set Data Clean up

- Data was pretty clean with no null values and all numerical data which fits well into any machine learning model
- Of the 12 variables- 10 are numerical and 2 categorical (4 wilderness types and 40 different Soil Types) which were both one hot encoded and that increased the dimensions to 56
- A completely balanced training set with equal values of the 7 cover types
- Careful analysis of the variables which were not one hot encoded showed that outliers were trivial and to some extent were helping the model so we decided to keep them

# Features Engineered

- New Columns created based on Hillshades
- Euclidean distance created for Horizontal_Distance_To_Hydrology and Vertical_Distance_To_Hydrology since they are highly correlated
- Merging of some of the Soil types since these were represented in more than 50% of the Cover Types
- Created new columns using elevation and distances
- Converted Aspect into a categorical variable and one hot encoded based on groups

# Models

- We used 2 models with Randomised Search Cross Validation
- Training set was split into 2 with 15% being the test set
- Random Forest Classifier and Extra trees Classifier were used:
  - Accuracy Score for Randomforest- 0.88
  - Accuracy Score for Extra Trees- 0.89
- **Our chosen model - Extra Trees Classifier**

# Conclusion & learnings

We managed to get a score of 78.8% on Kaggle with all the feature engineering done. Our model was not able to fully differentiate between Classes and 1 and 2  and also between 3 and 6. In future we would like to improve upon this drawback.

Also, it was challenging to predict half a million values by training on 15k.

# Final Submission on Kaggle with a score of 78.8%

| Submission and Description | Private Score | Public Score | Use for Final Score |
| --- | --- | --- | --- |
| result.csv<br>just now by Zhizheng Wang<br>add submission details | 0.78735 | 0.78735 | ☐ |

# Thank You