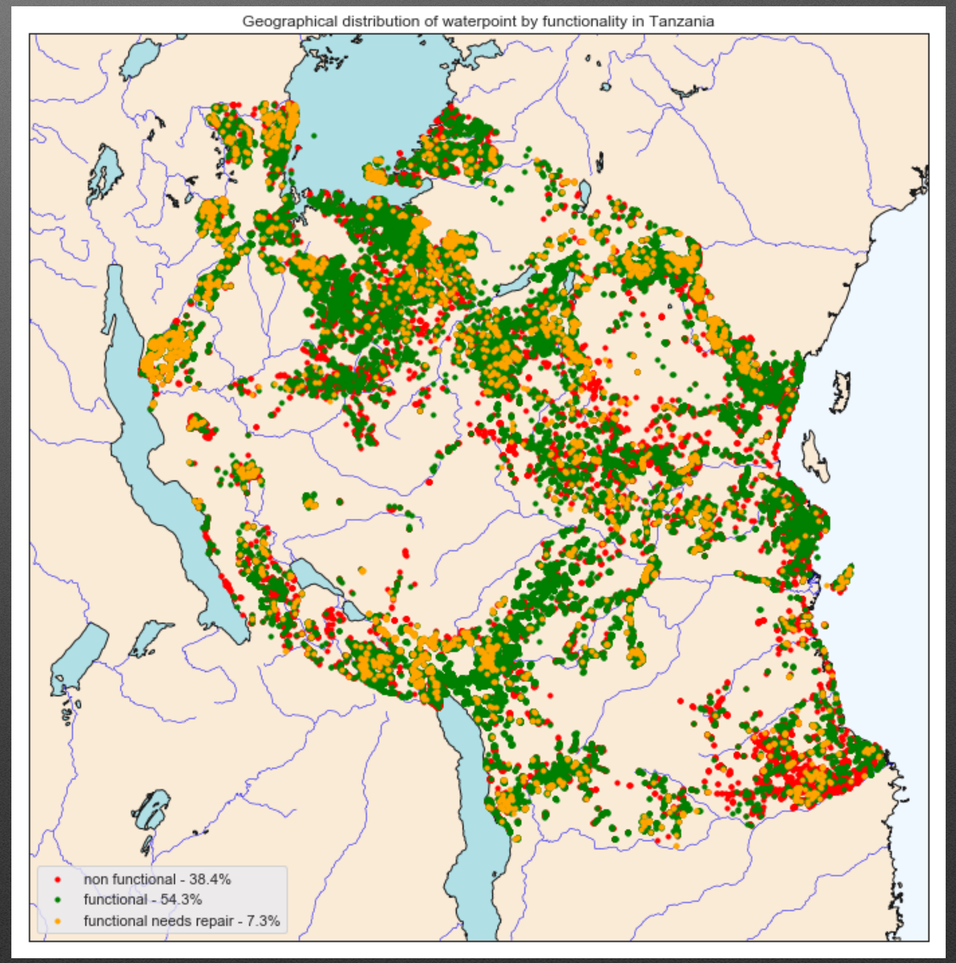
Pump It Up: Data Mining the Water Table

Flatiron Phase 3 Project
Joshua Blumer

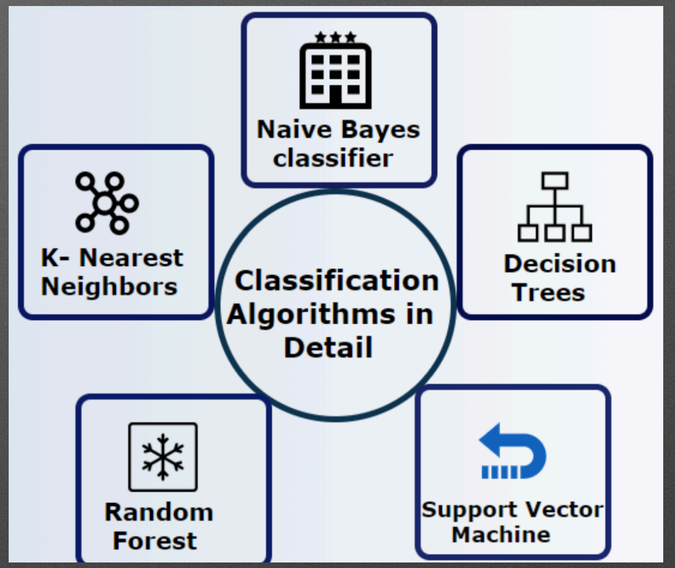
Introduction

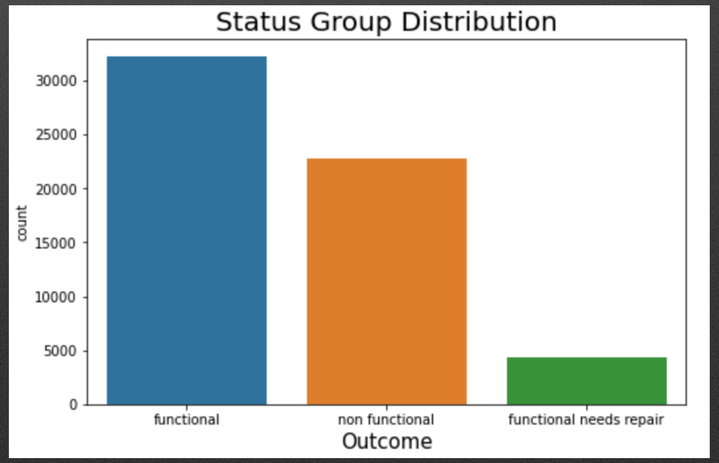
- An exploratory data analysis of the ‘Pump It Up: Data Mining the Water Table’ dataset provided by Driven Data
- Goal: For this project I am participating in an active classification competition. I will be modeling the dataset using classification techniques with the objective of obtaining as high of an accuracy score as possible



Methodology

- Preprocess data for multi-class classification models
- Explore different models
- Iterate through range of hyper-parameters to optimize greatest accuracy





Data Cleaning

- Clean and organize missing/null values
- Assess feature unique value counts and distributions
- Group features with manageable value counts by correcting syntax errors
- Remove features with excessive value counts to reduce dimensionality
- Remove features with low variance in target variable distribution

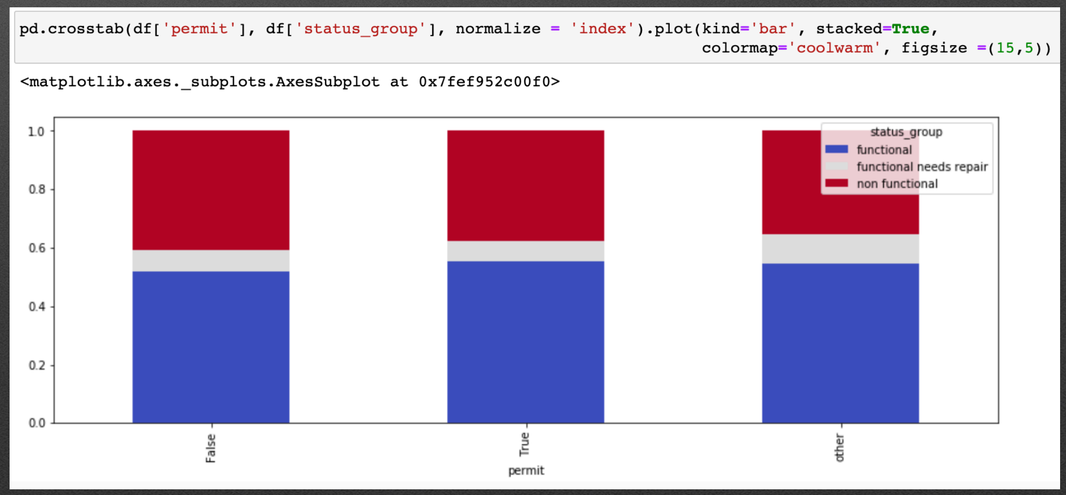
```
df_values['subvillage'].value_counts(normalize = True)
```

```
Madukani          0.008606
Shuleni           0.008572
Majengo           0.008504
Kati              0.006319
Mtakuja           0.004438
```

```
Nyaholo           0.000017
Isundama          0.000017
Masimila          0.000017
Mseza Mkulu       0.000017
Buzanaki          0.000017
Name: subvillage, Length: 19287, dtype: float64
```

```
pd.crosstab(df['permit'], df['status_group'], normalize = 'index').plot(kind='bar', stacked=True,
                                                                        colormap='coolwarm', figsize =(15,5))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fef952c00f0>
```



Preprocessing

- Check shape and arrangement of training and testing feature variables to ensure proper alignment
- One-hot encode categorical features
- Min-max scale separate instance of data to experiment and compare model outcomes

```
for col in processed_train.columns:
    print(col)
```

```
basin_Lake Nyasa
basin_Lake Rukwa
basin_Lake Tanganyika
basin_Lake Victoria
basin_Pangani
basin_Rufiji
basin_Ruvuma / Southern Coast
basin_Wami / Ruvu
region_Dar es Salaam
region_Dodoma
region_Iringa
region_Kagera
region_Kigoma
region_Kilimanjaro
region_Lindi
region_Manyara
region_Mara
region_Mbeya
region_Morogoro
```

```
for col in processed_test.columns:
    print(col)
```

```
basin_Lake Nyasa
basin_Lake Rukwa
basin_Lake Tanganyika
basin_Lake Victoria
basin_Pangani
```

Modeling

- Model data iteratively using multiple classification techniques
- Begin with naive models (default settings)
- Choose strongest candidates to perform hyper-parameter grid searches and cross validation
- Experiment with returned grid search hyper-parameters on various combinations of feature selections

```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=1, gamma=0, gpu_id=-1,
              importance_type='gain', interaction_constraints='',
              learning_rate=0.1, max_delta_step=0, max_depth=9,
              min_child_weight=1, missing=None, monotone_constraints='()',
              n_estimators=400, n_jobs=-1, nthread=4, num_parallel_tree=1,
              objective='multi:softprob', random_state=0, reg_alpha=0,
              reg_lambda=1, scale_pos_weight=1, seed=0, silent=True,
              subsample=0.5, tree_method='exact', validate_parameters=1,
              verbosity=None)
```

```
# Predict on training and test sets with classifier 2

test_preds2 = clf2.predict(X_test)
```

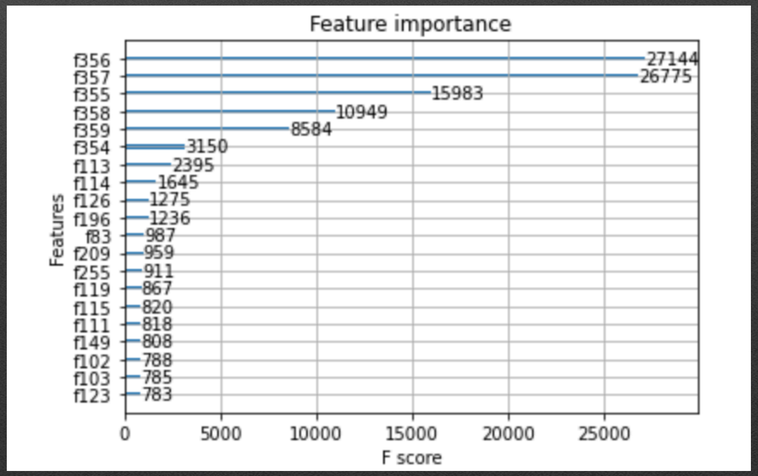
```
# Accuracy of training and test sets for classifier 2

training_accuracy = clf2.score(X_train, y)

print('Training Accuracy: {:.4}%'.format(training_accuracy * 100))

# 250 Estimators: 88.05%
# 400 Estimators: 90.94%
```

```
Training Accuracy: 90.94%
```



Conclusion

- After experimenting with many iterations of preprocessed data combinations and models I achieved a score of 80.93% accuracy placing me within the top 19% of all submissions for the competition
- With more time to work on this dataset I would further analyze feature importances and attempt optimizing a partition of the data for cat boost modeling

Submissions

BEST	CURRENT RANK	# COMPETITORS
0.8093	2299	12335