Telstra Network Disruptions

By Jerome Benton, Senior Data Scientist at Carfeine

What factors have the most impact on service disruptions?

Metric

Logarithmic Loss

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij})$$

Submission Format

ID	sev0	sev1	sev2
12345	predict_proba	predict_proba	predict_proba

The Data

Provided by

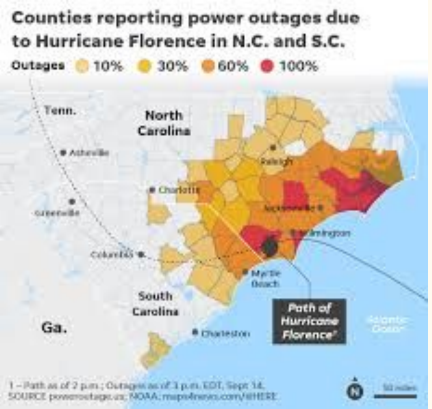
Kaggle

- Train - training set
- Test - testing set
- Sample - submission
- Event_type - encoded events
- Log_feature - encoded feature and volume
- Resource_type - encoded resource
- Severity_type - encoded severity

EDA &
Feature Engineering

In the Real world

1. Location
2. Weather
3. Time (of day, month, year)
4. Maintenance
5. Power failure
6. Natural disaster
7. Demand fluctuation



Counties reporting power outages due to Hurricane Florence in N.C. and S.C.

1 – Path as of 2 p.m.; Outages as of 3 p.m. EDT, Sept. 14.
SOURCE poweroutage.us; NOAA; maps4news.com/HERE

In the Kaggle World

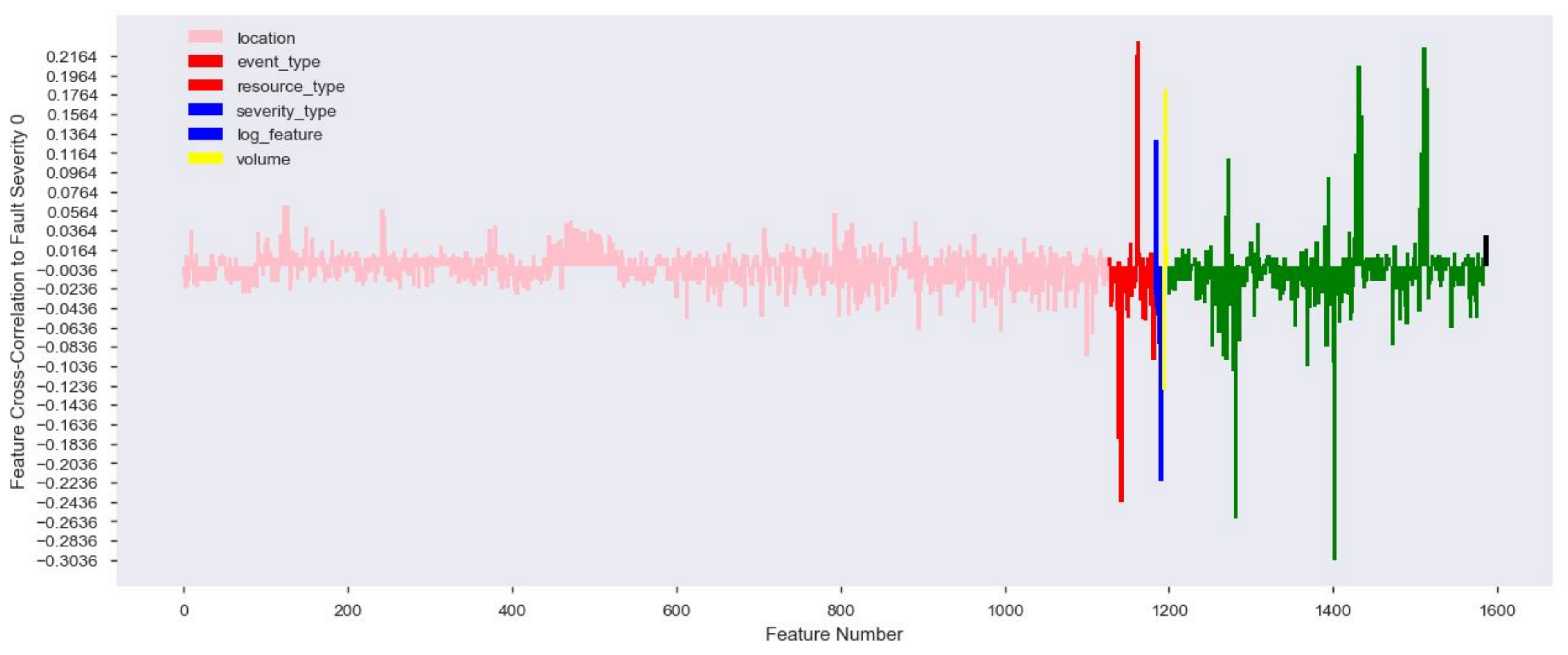
Cryptic Names and Encoded Values

1) Locations - order that suggests multiple encoding
2) Events - encoded, correlation on many, but not all
3) Log types - encoded syslog type error levels
4) Volumes - encoded, correlations to log entries
5) Resources - appears to be related to network elements
6) Severity - encoded priority, don't confuse with fault types
7) Fault Severity - relative weight of the network disruption

Dirty Data

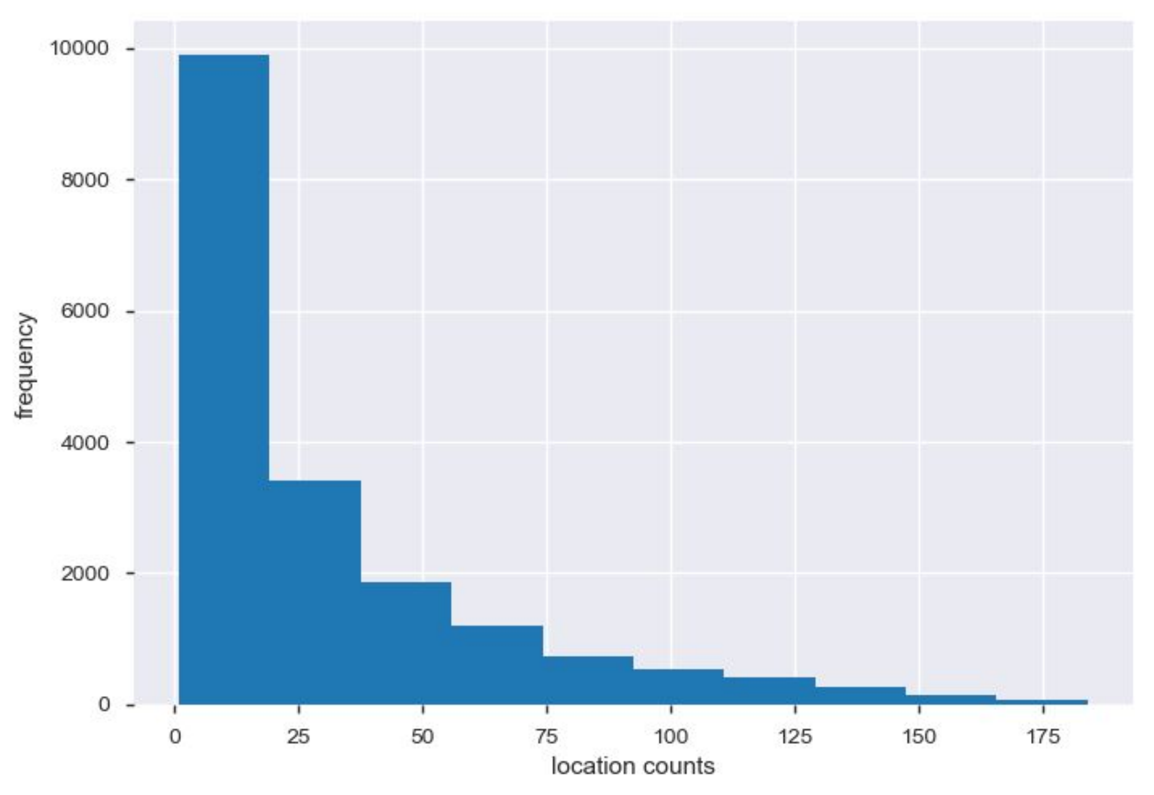
1) Corrupt files - two corrupt files
2) Data types - all over the place
3) Missing data - luckily, MCAR
4) Imbalanced classes - can skew results if not prepared
5) Feature Engineering - required if you want to score above 50%

Correlation Analysis by Feature



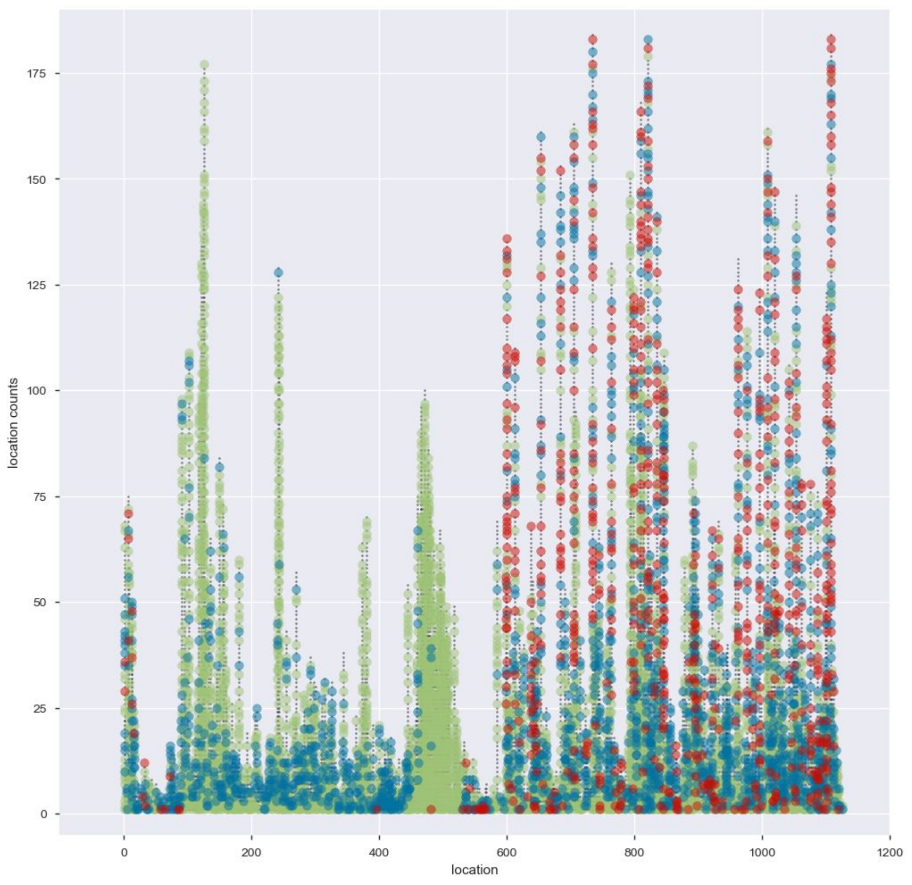
Location Counts

Aggregating locations in this histogram illustrates the frequency of incidents at locations



Location, location, location

Not only was location encoded with a temporal variable, this scatter plot illustrates the relationship between location and fault severity.



Model & Evaluation

First Try

XGBClassifier
Without FE
Grid Search Tuned

Model log-loss: 0.365875714755

Kaggle Private log-loss: 0.58970

Private Leaderboard rank: 572

Percent Rank: 41.3%

Best Try

XGBClassifier
With Extensive FE
Grid Search Tuned

Model log-loss: 0.291117886791

Kaggle Private log-loss: 0.45170

Private Leaderboard rank: 74

Percent Rank: 92.4%

Conclusion

All the hard work of feature engineering paid off with a Kaggle private leaderboard jump from 572 to 74, from 41% to 92 %.

Questions?