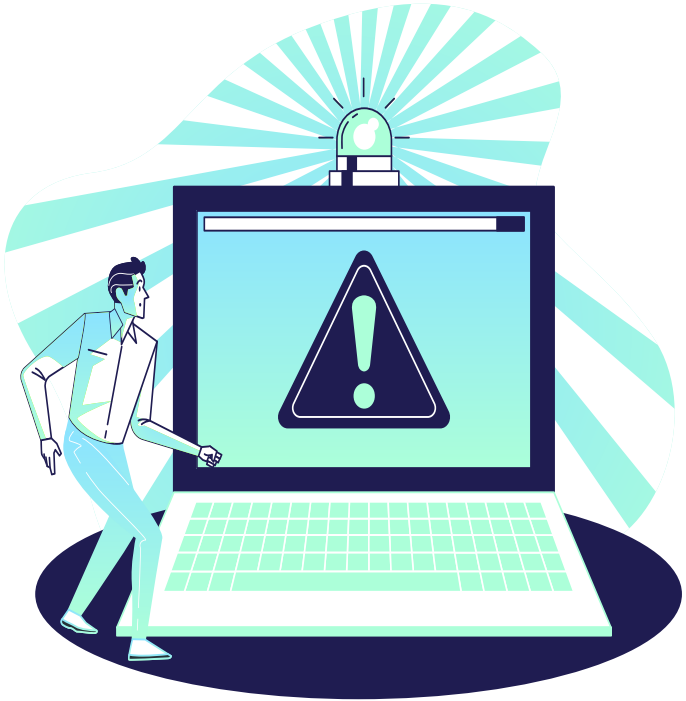
Data Mining and Machine Learning



TESLA

Inessa Khaneeva, Jakob Kampik, Mariem Guebibia

Intro explanation approach

Task

“Predict which tweets are about a real disaster and which are not”.

Our approach:

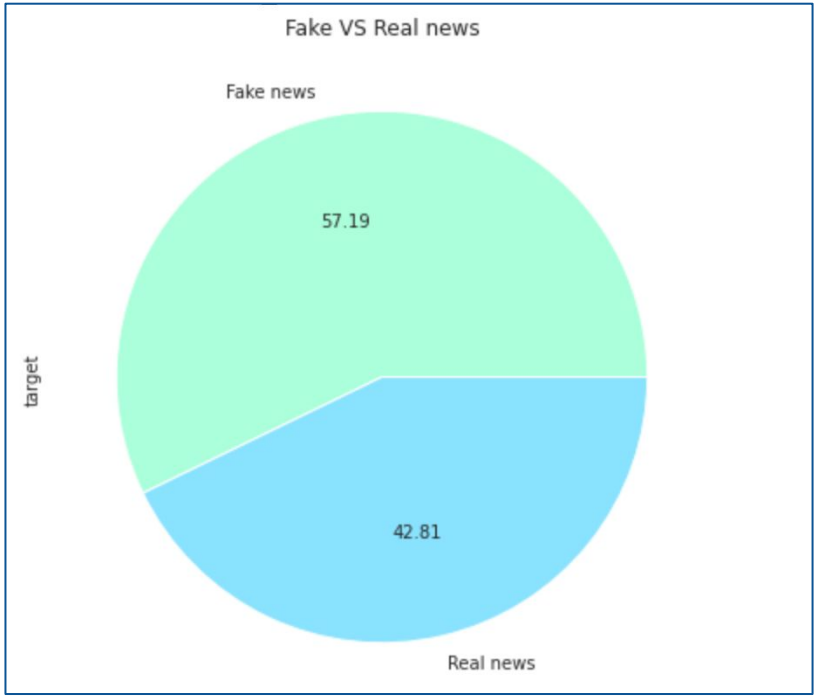
- **First EDA.**
- **Then, work on iterations to improve accuracy.**

EDA 1

Training data
- 6471 observations
- 5 features
- “Id”, “keyword”, “location”, “text”, “target”
- Types: Int 64, object

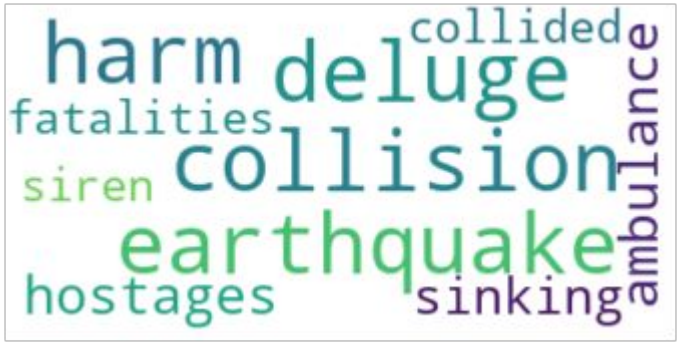
Base rate: 0.5719

→ not hard to find
but hard to tell apart

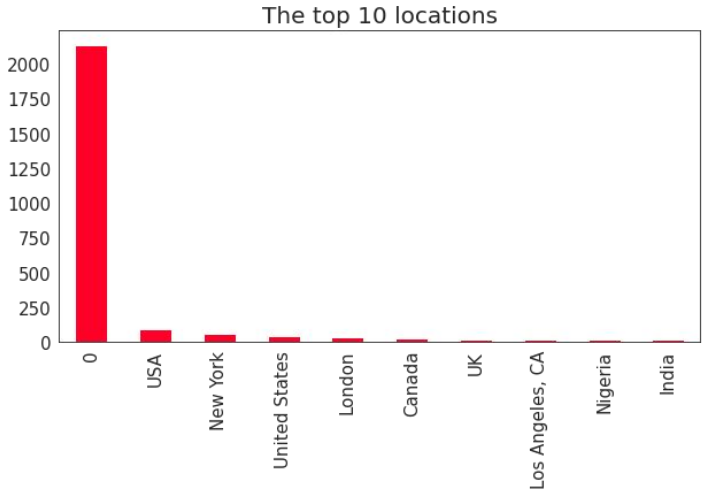


EDA 2

Most common keywords

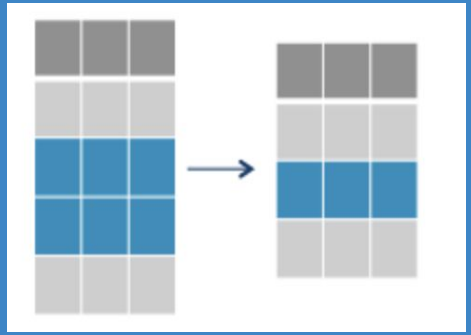


Most common locations



1st Iteration Data Cleaning

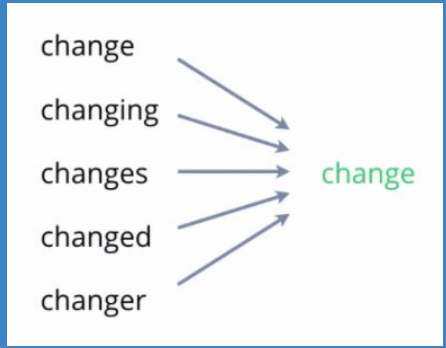
Duplicate tweets



Stopwords & punctuation

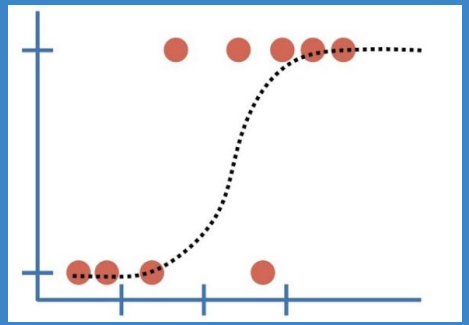
, . - ! ? ‘ “

Lemmatization



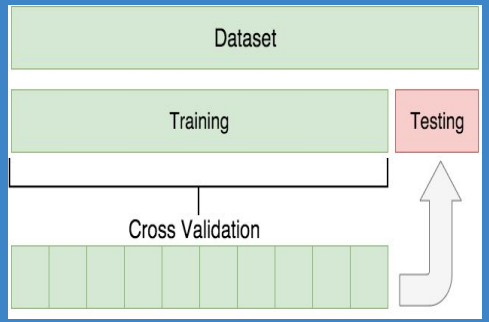
1st Iteration Prediction

Logistic Regression



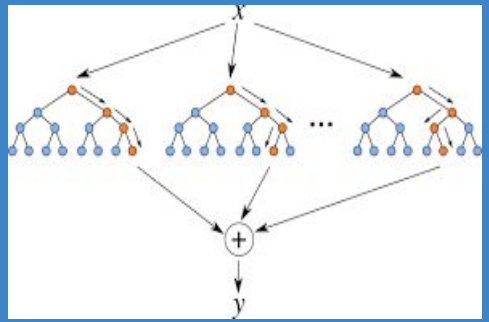
Score: 0.808
Score: 0.810

Cross Validation



Score: 0.790

Random Forest



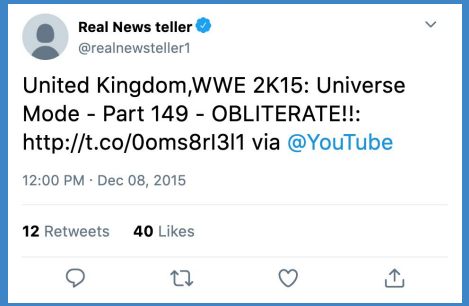
Score: 0.797

2nd Iteration more Data Cleaning

HTML Chunks

https*\S+
@\S
#\S+
\'\w+
\w*\d+\w*
\s{2,}

Advanced Cleaning with PorterStemmer

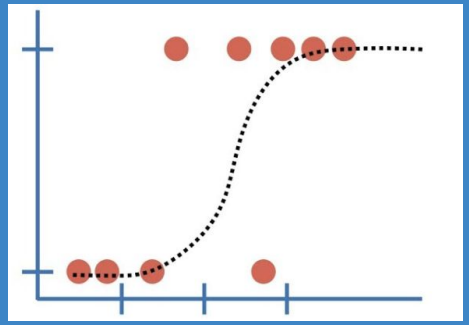


2015, 2k15 > same meaning

```
tweet = re.sub(r"2k15", "2015", tweet)
```

2nd Iteration Prediction

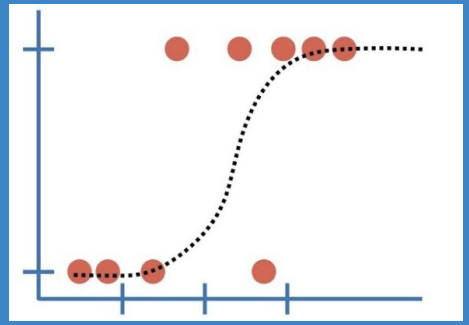
Logistic Regression 2nd iteration



Score without CV: 0.796

Score with CV: 0.788

Logistic Regression from 1st iteration



Score: 0.811 🎉 - AICROWD

Our Highest Accuracy 🎉

Score: 0.811

Classifier:
Logistic regression without cross-validation

Determinants:
- **Simple cleaning**
- **Simple classifier**

3rd Iteration: Improving classifiers

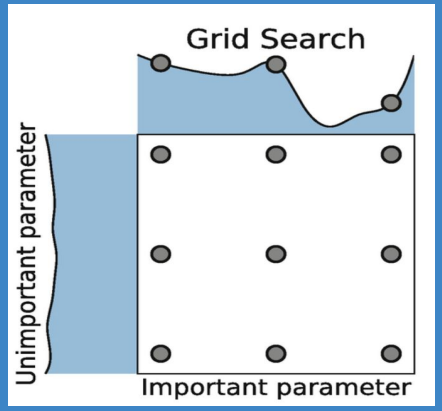
Goal:

- **Improve accuracy further**

Considerations:

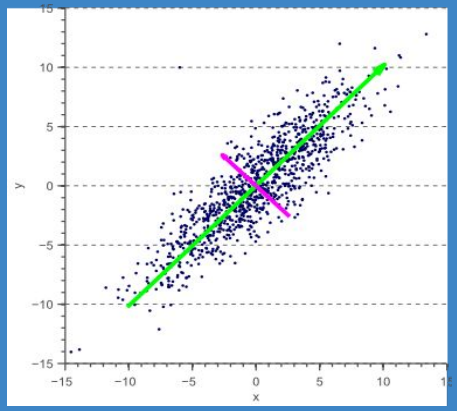
- **More sophisticated classifiers**
- **Dimensionality reduction**

Logistic Regression with Grid Search



Score: 0.793

Logistic Regression with PCA



Score: 0.792

3rd iteration: More classifiers

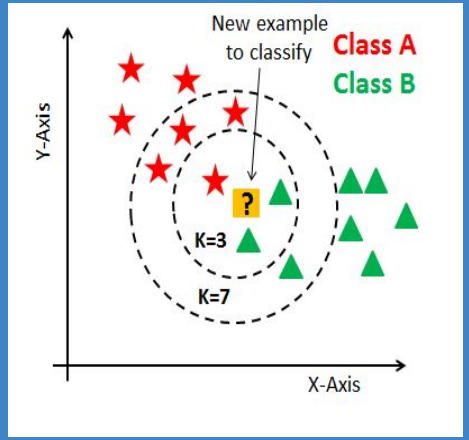
Goal:

- **Improve accuracy further**

Considerations:

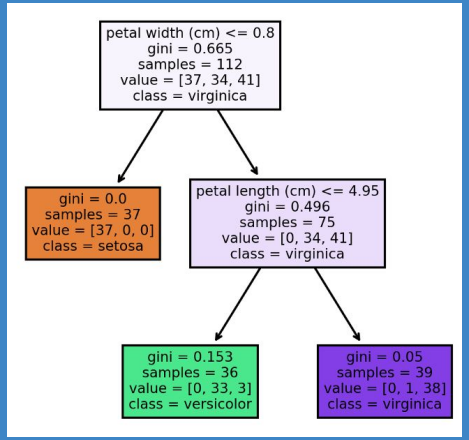
- **Less sophisticated classifiers**
- **Better generalization**

kNN



Score: 0.660

Decision Trees



Score: 0.715

Conclusion

Reflection:

More sophisticated classifiers did not improve the code.

Logistic regression without Cross validation delivered the highest accuracy.

Table of Images

Duplicate Tweets: https://www.datanovia.com/en/lessons/identify-and-remove-duplicate-data-in-r/
Lemmatization: https://medium.com/swlh/introduction-to-stemming-vs-lemmatization-nlp-8c69eb43ecfe
Logistic Regression: https://www.youtube.com/watch?v=yIYKR4sgzI8
Cross Validation: https://medium.com/@rj322198/cross-validation-in-machine-learning-c677653ea475
Random Forest: https://levelup.gitconnected.com/random-forest-regression-209c0f354c84
TD-IDF: Lecture Slides by Prof. Vlachos
Grid Search:
https://medium.com/@cjl2fv/an-intro-to-hyper-parameter-optimization-using-grid-search-and-random-search-d73b9834ca0a
PCA: https://medium.com/analytics-vidhya/a-deep-dive-into-principal-component-analysis-pca-4e8a6d5a6386
kNN: https://www.kdnuggets.com/2020/11/most-popular-distance-metrics-knn.html
Decision trees: https://www.kdnuggets.com/2020/04/visualizing-decision-trees-python.html