Prediction of volcanic eruption.

Group No. 11
Rushabh Nalin Mehta
Mentor: Gaurav Kumar

Agenda

- Necessity of the project
- Problem Statement
- Methodology
- Data Analytics & Model Training
- Results
- Conclusion
- Acknowledgements

Necessity of the project

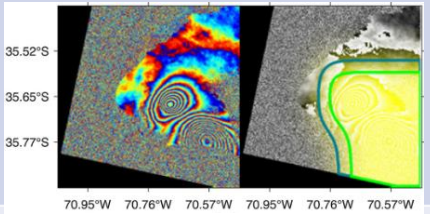
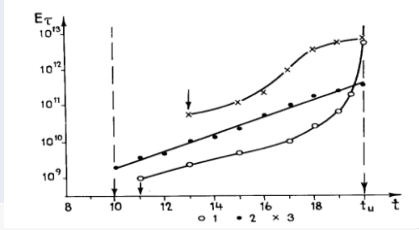
- Following industries have interests in volcanic active areas.
  - Mining -> Mining towns
  - Geothermal Energy
  - Tourism -> Hotels, restaurants and spas
  - Agriculture -> Agriculture settlements
- Eruption will lead to destruction of life and property if not evacuated
- Timely evacuation requires a robust and reliable prediction system

Problem statement

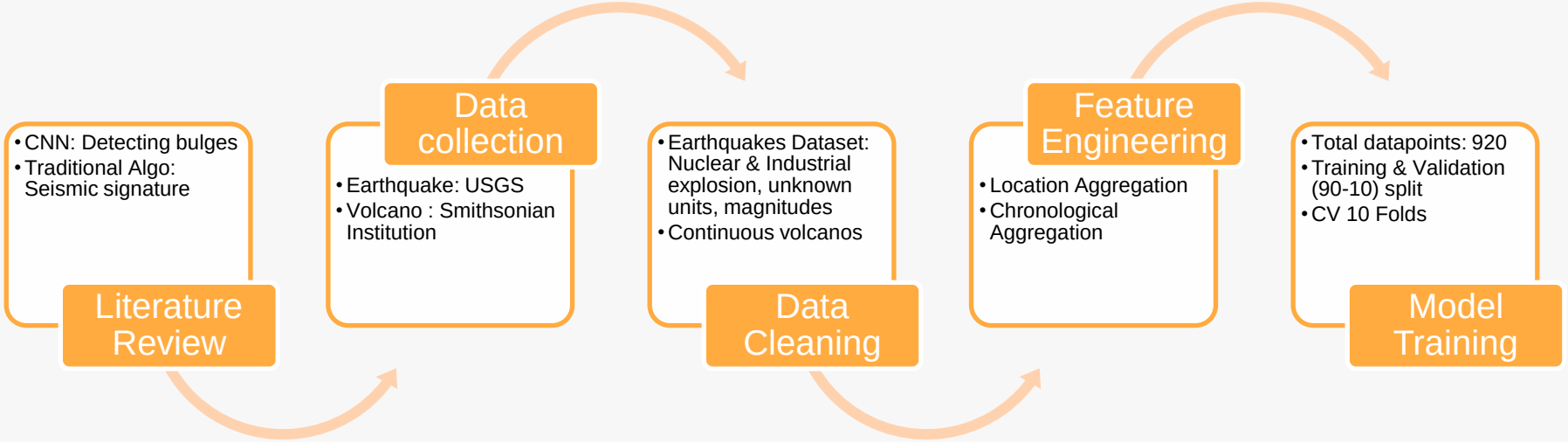
- Develop a machine learning model to predict if there will be a volcanic eruption in the next 6 months.

Literature Review

- We went through plenty of papers and realized that most of them referred to seismic signature of eruptive behavior of the volcanos and hence we decided to use seismic data to predict the earthquake

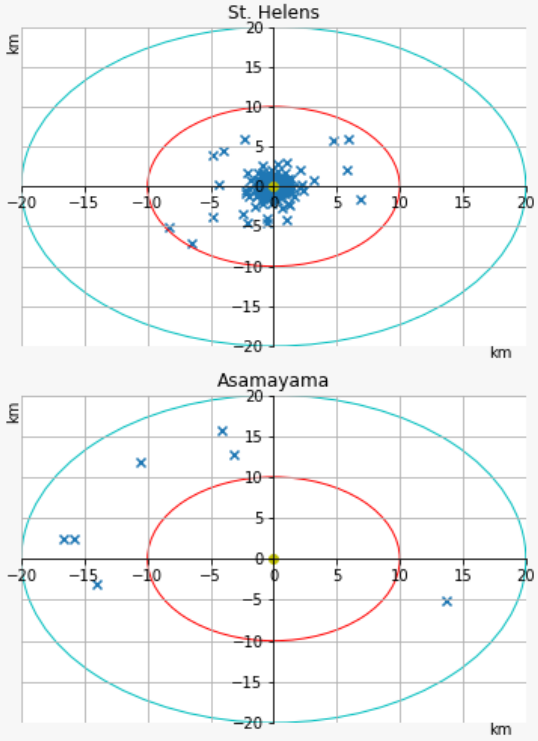
Paper Name	Comment	Interesting images
Application of Machine Learning to Classification of Volcanic Deformation in Routinely Generated InSAR Data	Which used satellite imagery to detect surface swelling which is an early sign of volcano eruption. CNN was applied to turn this into a predictive model. Different volcano types were included in the analysis	
The prediction of large explosions of andesitic volcanoes by P.I. Tokarev	Data is transformed into energy release over the period and then it is used as an estimator for predicting the volcano	

Method


Literature Review
• CNN: Detecting bulges
• Traditional Algo: Seismic signature
Data collection
• Earthquake: USGS
• Volcano : Smithsonian Institution
Data Cleaning
• Earthquakes Dataset: Nuclear & Industrial explosion, unknown units, magnitudes
• Continuous volcanos
Feature Engineering
• Location Aggregation
• Chronological Aggregation
Model Training
• Total datapoints: 920
• Training & Validation (90-10) split
• CV 10 Folds

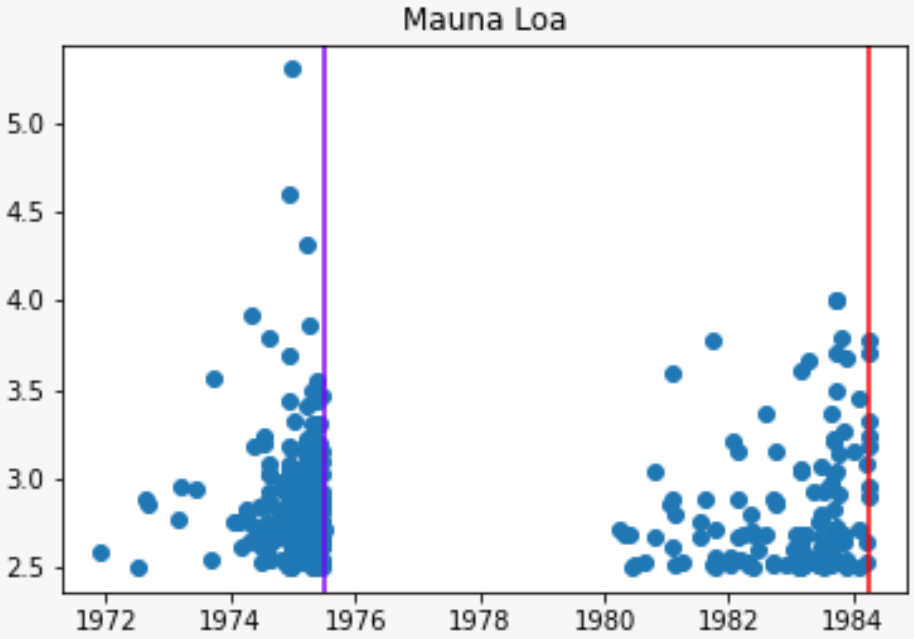
Data Aggregation

- Aggregation by location
  - All earthquakes in 10 km radius
  - Minimum 10 entries subjected to 20 km hard radius cut
  - This is done in accordance with paper by P.I. Tokarev. It states that earthquakes that have the epicenter near the volcano have higher correlations with eruptions



Data Aggregation (Conti.)

- Aggregation by Chronological order
  - All earthquake that occur between the start date of current volcano and end date of the last volcano are considered
  - If the said period is longer than 4 years, earthquakes of the last four years are considered.



Model training

- We split the dataset into 90% training and 10% validation set.
- We further use 10-fold cross validation with 70-30 split
- Logistic regression and Xgboost using scikit-learn
- ML library pycaret used for train and evaluate on the following traditional machine learning algorithms
    - K Neighbours classifier, Ridge Classifier, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Logistic Regression, Ada Boost Classifier, Random Forest Classifier, Gradient Boosting Classifier, CatBoost Classifier, SVM - Linear Kernal, Extra Trees Classifier, Light Gradient Boosting Machine, Extreme Gradient Boosting, Decision Tree Classifier, Naive Bayes

Results interpretation

Results from Auto ML library Pycaret

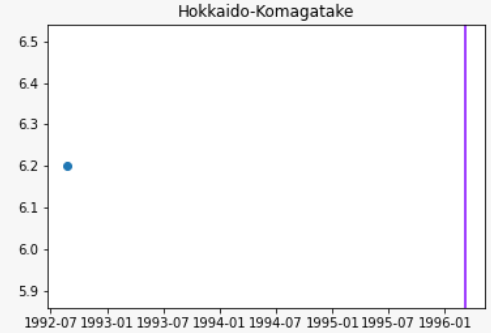
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
knn	K Neighbors Classifier	0.8311	0.8950	0.8476	0.7884	0.8129	0.6593	0.6672	0.2940
ridge	Ridge Classifier	0.8200	0.0000	0.9642	0.7216	0.8241	0.6476	0.6800	0.0100
lda	Linear Discriminant Analysis	0.8200	0.8843	0.9537	0.7248	0.8227	0.6469	0.6744	0.0080
qda	Quadratic Discriminant Analysis	0.8178	0.8340	0.9684	0.7164	0.8223	0.6433	0.6786	0.0090
lr	Logistic Regression	0.8111	0.8961	0.8937	0.7334	0.8028	0.6245	0.6422	0.6110
ada	Ada Boost Classifier	0.8044	0.8990	0.8168	0.7593	0.7819	0.6052	0.6145	0.0260
rf	Random Forest Classifier	0.7978	0.8958	0.7821	0.7602	0.7654	0.5882	0.5957	0.0630
gbc	Gradient Boosting Classifier	0.7956	0.8845	0.8124	0.7412	0.7694	0.5866	0.5970	0.0230
catboost	CatBoost Classifier	0.7956	0.8917	0.8224	0.7390	0.7733	0.5881	0.5989	0.4850
svm	SVM - Linear Kernel	0.7889	0.0000	0.8234	0.7379	0.7731	0.5776	0.5874	0.0080
et	Extra Trees Classifier	0.7867	0.8754	0.7653	0.7483	0.7511	0.5644	0.5704	0.0610
lightgbm	Light Gradient Boosting Machine	0.7844	0.9000	0.7511	0.7492	0.7442	0.5588	0.5661	0.0400
xgboost	Extreme Gradient Boosting	0.7800	0.8960	0.7308	0.7521	0.7329	0.5477	0.5564	0.1210
dt	Decision Tree Classifier	0.7533	0.7773	0.6595	0.7435	0.6891	0.4889	0.4974	0.0070
nb	Naive Bayes	0.7289	0.8949	0.4503	0.8463	0.5429	0.4078	0.4597	0.0080

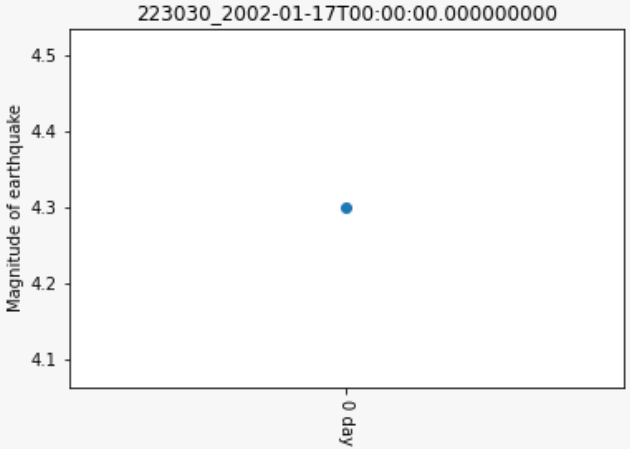
Results from Scikit Learn Implementation

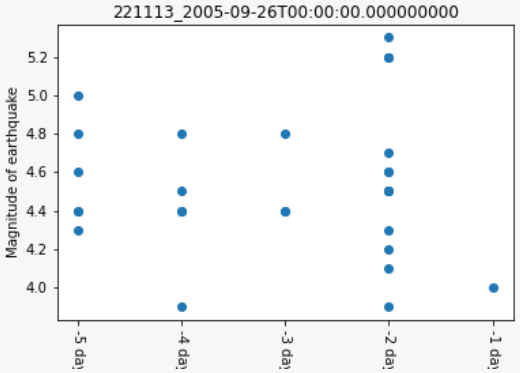
- Logistic Regression
  - Classification accuracy: 0.84
  - Recall: 1
  - Prec.: 0.78
  - F1: 0.87
- Xgboost
  - Classification accuracy: 0.84
  - Recall: 0.95
  - Prec.: 0.86
  - F1: 0.91

Conclusion

- Auto ML library like pycaret is an amazing way to compare different traditional algorithms
- Seismic activity can’t be the only feature we consider







Acknowledgements

- Mentor
  - Appreciates and recognises the work done by the team members on the project.
  - Opportunity given by Techlabs Aachen e.V. to manage a team of data engineers in complex project is appreciated
- Team Members
  - The knowledge related to data science is significantly enhanced by completing the track as well as the practical project
  - The well-organized event, reliable assistant and passionate team members of Techlabs Aachen e.V. was one of the factors that help us to complete the project.

Open for Questions!

Thank you for your attention

Rushabh Nalin Mehta (rushabh.mehta@rwth-aachen.de)

Mentor : Gaurav Kumar (gauravkumarjha1195@gmail.com)