# Earthquake Time Prediction using CatBoost and SVR

**Sahaya Sakila, Sanyam Garg, Tanay Yeole , Hrithik Yadav**

***Abstract*: *Seismic tremors everywhere throughout the globe have been a noteworthy reason for decimation and death toll and property. The following context expects to recognize earthquakes at a beginning time utilizing AI. This will help individuals and salvage groups to make their errand simpler. The information in this manner comprises of these seismic acoustic signals and the time of failure. The model is then prepared utilizing the CatBoost model and the utilization of Support Vector Machines. This will help foresee the time at which a Seismic tremor may happen. CatBoost Regression Algorithm gives a Mean Absolute Error of about 1.860. The Cross Validation (CV) Score for the Support Vector Machine (SVM) approach is -2.1651. The datasets metrics are not reliable on any outer parameter in this manner the variety of exactness is constrained, and high accuracy is accomplished.***

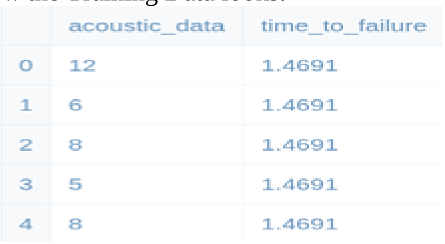
***Keywords* : *CatBoosting, Support Vector Regression, Acoustic signals, Earthquake, Exploratory Data Analysis, Skew, RMS, Peak.***

## I. INTRODUCTION

This content is exhibited for the examination of seismic-acoustic signals, which permits the utilization of tremor in the sign as a transporter of data toward the beginning of typical seismic procedures seismic tremors[1]. So as to limit potential results, it is mandatory to discover early cautioning procedures to alarm the populace on frameworks created by an enormous number of observing stations, prepared with high exactness instruments fit for chronicle seismic action, have been sent around the outside of the most dynamic volcanoes everywhere throughout the world[2] .A solid seismic tremor might be trailed by numerous delayed repercussions. At the point when encompassing seismological stations distinguish those delayed repercussion signals, we could decide the area of consequential convulsions by breaking down the time that vertical (P) wave and even (S) wave come to the seismograph station[3] . Every seismogram can contain different sorts of seismic occasions, and now and again it is additionally conceivable to catch occasions of non-volcanic starting point, for example, lightning[2] .The data originates from a notable trial set-up used to ponder tremor material science. The acoustic information info sign is utilized to anticipate the time staying before the following research facility seismic tremor. The training data is a solitary, nonstop fragment of test information. The test information comprises of a folder containing numerous little segments. The information inside each test document is constant, however the test records don't represent to a ceaseless portion of the investigation; along these lines, the forecasts can't be accepted to pursue a similar customary example found in the preparation document. The model has an accuracy of 89.99% in A, P, and R. The model performs 95.99% authentic on the S metrical by using the SMO technique. 59.99% of the data from the dataset are chance to chosen for training while the other use 39.99% were used for testing, which is not a proper ratio to divide a dataset. The dataset used contains a lot of noise and is therefore hard to process. The model has a Mean Absolute Error (MAE) of about 1.860 for the Cat Boost Regression Algorithm. The Cross Validation (CV) Score for the Support Vector Machine (SVM) approach is -2.1651. The dataset metrics are not dependent on any external parameter therefore the variation of accuracy is limited, and high precision is achieved.

## II. SYSTEM ARCHITECTURE

Using a seismometer, seismic signals are mostly registered and these signals contain a lot of noise. In this case, we use signals recorded in a laboratory so we won't be worried with removing signals as we concentrate primarily on anticipating the Time of Failure or the time when an Earthquake happens. The first thing after obtaining the data is to split the data into Training and Testing Data.
This is how the Training Data looks:



| | acoustic_data | time_to_failure |
|---|---|---|
| 0 | 12 | 1.4691 |
| 1 | 6 | 1.4691 |
| 2 | 8 | 1.4691 |
| 3 | 5 | 1.4691 |
| 4 | 8 | 1.4691 |

**Fig 1: A snapshot of the first few training points**

Revised Manuscript Received on November 30, 2019.
* Correspondence Author

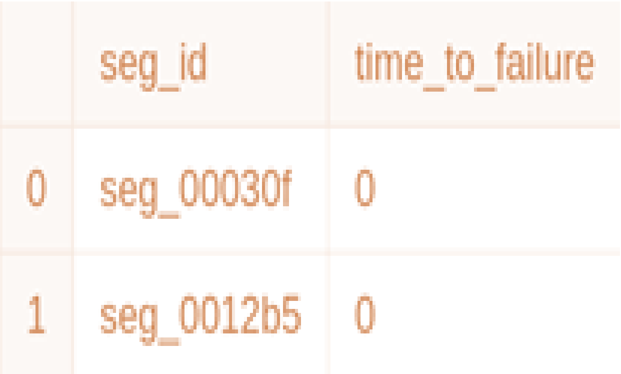
**Sahaya Sakila***, Assistant Professor, Computer Science And Engineering Department, in SRM Institute of Science and Technology, Ramapuram, Chennai, India.

**Sanyam Garg,** Pre-Final Year Student of B.Tech Computer Science And Engineering in SRM Institute of Science and Technology, Ramapuram, Chennai, India.

**Tanay Yeole** Pre-Final Year Student of B.Tech Computer Science And Engineering in SRM Institute of Science and Technology, Ramapuram, Chennai, India.

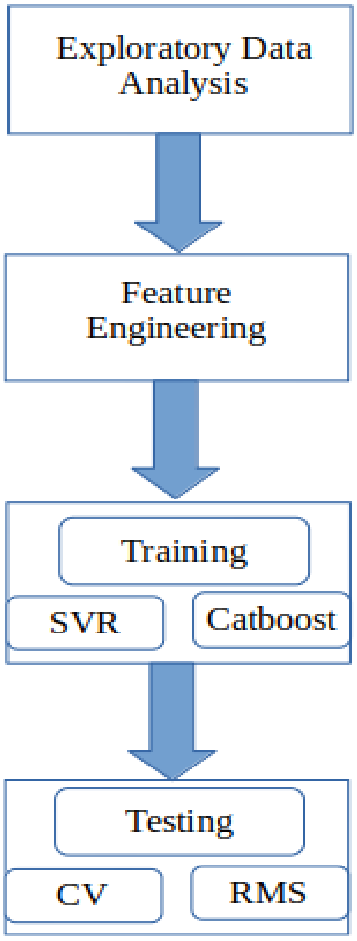
**Hrithik Yadav,** Pre-Final Year Student of B.Tech Computer Science And Engineering in SRM Institute of Science and Technology, Ramapuram, Chennai, India.

This is how the Testing Data looks like:



| | seg_id | time_to_failure |
|---|---|---|
| 0 | seg_00030f | 0 |
| 1 | seg_0012b5 | 0 |

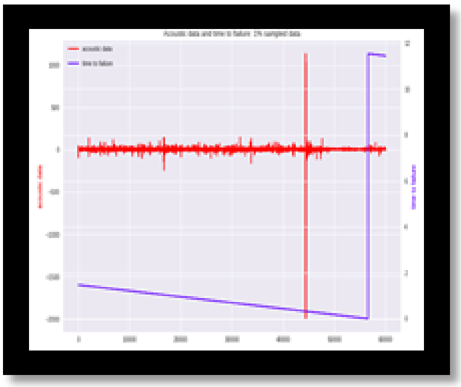
**Fig 2:** A snapshot of the first few Testing points The next step is to visualize the data after this. We use Exploratory Data Analysis (EDA). EDA is a significant step in making it possible for us to view the information. As the information is restricted to the seismic signal and the time of failure discussed, we use Feature Engineering to improve the amount of features in the dataset. This provides us space for information exploration and outcomes to be more precise. We then use the CatBoost Regression Model and the Support Vector Regression Model to make valuable predictions.



### 1.1. Exploratory Data Analysis:

The training data used here is stored here in the form of CSV files. The whole data has been extracted into a pandas data frame. This data will be further manipulated and 1% of this data i.e. 100 Data Points will be taken into account and then been visualized.

The following result is then obtained



The Red Line here is the signal input or the acoustic data. The Blue Line is the moment of failure or the moment of the occurrence of an earthquake. Here it can be noted that when a sudden spike occurs in the Seismic Signal (the Red Line), the Blue Line goes up. This informs us that after the spike the Earthquake happens. The Similar Thing is Observed for 100% of the Data. So we have got a pattern here, which was possible due to this step of Exploratory Data Analysis.

### 2.2 Feature Engineering:

We can see from the Figure 1 and Figure 2 that there are only two columns. One is the Feature and the latter is the label. In this Feature Engineering Step we are Going to increase the number of features. We will be adding about 8 to 9 features so as to make our predictions more accurate.

#### 2.2.1 Mean :

The first will be the mean, the average value of the whole dataset.

$$\underline{a} = (\sum g a)/b$$

$\underline{a}$ = Mean

g = frequency of each class

a = mid-interval value of each class

b = total frequency

#### 2.2.2 Standard Deviation:

The next is Standard Deviation, which tells how spread our dataset is:

$$s_e = (\sqrt{\sum_{i=1}^{b} b(\,a_i - \underline{a}\,)^{2}}\,)/(b-1)$$

b = Number of Data Points

$\underline{a}$ = The mean of a $a_i$

$a_i$ = Each of the values of the Data

#### 2.2.3 Kurtosis:

Kurtosis is a calibrate of whether data are heavily-tail or lightly-tail relational to regular allocation.

Datasets with high Kurtosis will in general have substantial tails, or anomalies. Informational indexes with low kurtosis will in general have light tails, or absence of exceptions.

$$Kurtosis = ((\sum_{i=1}^{b}(a_i - \underline{a})) / (B)) / (D^4)$$

Here,

a = The mean of $a_i$

$\underline{a}$ = Each of the values of the Data
D = Standard Deviation
B = The sample size

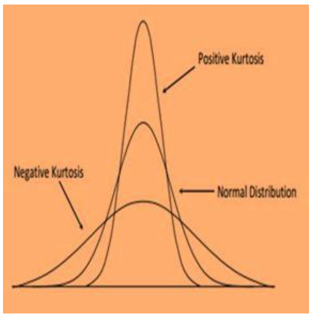


**Fig 5: A visual Representation of** Positive and Negative Kurtosis with the Normal Distribution.

### 2.2.4 Skew:

Skewness is asymmetry in a factual dispersion in which either to one side or to the privilege the bend seems twisted or slanted. How much an appropriation contrasts from a typical conveyance can be quantified. The skewness worth can be sure or negative, or indistinct.

When allocation is curled to the left handed side, the tail on the curve left-hand side is lengthy than the tail on the right-handed side, and the mean is smaller than the mode. This circumstance is additionally called negative skewness. The tail on the correct side of the bend is lengthy than the tail on the left handed side when a dissemination is slanted to one side, and the middle is higher than the mode. Positive skewness is likewise called this circumstance.

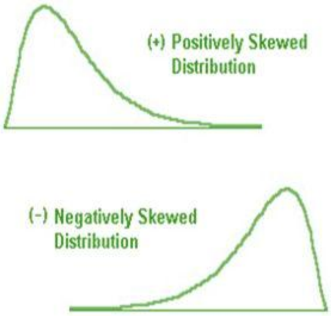


**Fig 6: A visual Representation of** Positive and Negative Skewness.

### 2.2.5 Quantile:

Quantiles are cut focuses that chasm the assortment of a likelihood dissemination into steady interims with equivalent probabilities, or likewise isolate the perceptions into an example.

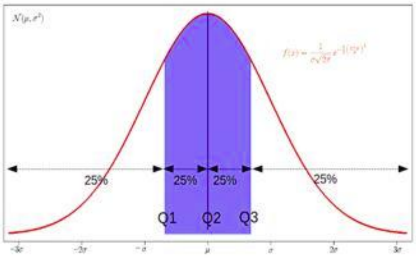


**Fig 7: A visual Representation of Quantiles**

### 2.2 Training & Testing:

This is the component where the data set we have is trained and the two major models used in the project are implemented.

Let's begin with CatBoost.

### 2.3.1 CatBoost:

CatBoost is an ensemble Machine Learning Method by Yandex ( a Russian Search Engine)."CatBoost" name originates from two words "Classification" and "Boosting". This turns out as a much better technique than Random Forest and other Gradient Boosting Methods.

There are two reasons for this :

● It yields cutting edge results without broad information preparing normally required by other AI strategies

● Gives ground-breaking out-of-the-case support for the more expressive information designs that go with numerous business issues.

This is the reason we use CatBoost as our method of gradient boosting. This is basically a Decision Trees ensemble.

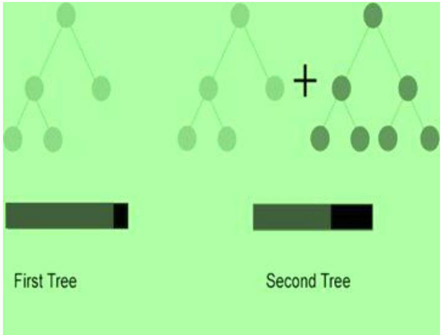


**Fig 8: An Ensemble of Decision Trees**

The calculation become familiar with the main tree to diminish the preparation botch on the primary cycle appeared in figure 8 on the left picture. This model more often than not has a noteworthy mistake; it is anything but a smart thought to fabricate huge trees in boosting since they overfit the information.

The image on the privilege in Figure 8 shows the second emphasis wherein the calculation learns another tree to diminish the primary tree's mix-up.

The calculation rehashes this technique until it manufactures a not too bad quality model.

The steps we follow are:

1.We first model information with basic models and investigate information for mistakes.

2.These blunders mean information indicates that are troublesome fit by a basic model.

3.At that point for later models, we especially center around those difficult to fit information to get them right.

4. At last, we consolidate every one of the indicators by giving a few loads to every indicator .

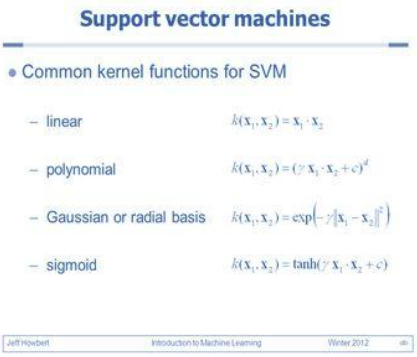
We get the Mean Absolute Error (MAE) as 1.74357544138688 after testing or modeling with the test information.The Mean Absolute Error is a loss function for Regression Models in particular. It is the sum of our target's absolute variations and predicted variables. Therefore, in a set of projections, it measures the average magnitude of mistakes without considering their directions.It's Formula goes as follows:

$$MAE = (\sum_{i=1}^{b} \quad (|s_i - s_i^e|))/b$$

### 2.3.2 Support Vector Regression:

Given two items, the bit yields some comparability score. The articles can be anything beginning from two whole numbers, two genuine esteemed vectors, trees whatever gave that the part capacity realizes how to think about them.

Another fascinating piece models is Gaussian portion. Given two vectors, the likeness will decrease with the span of σ. The separate two articles are "reweighted" by this span parameter. The accomplishment of learning with portions (once more, in any event for SVMs), unequivocally relies upon the decision of part. We found this Gaussian or Radial Basis Function Kernel very effective in our Project.



Support vector machines

- Common kernel functions for SVM
  - linear $k(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1 \cdot \mathbf{x}_2$
  - polynomial $k(\mathbf{x}_1, \mathbf{x}_2) = (\gamma\, \mathbf{x}_1 \cdot \mathbf{x}_2 + c)^d$
  - Gaussian or radial basis $k(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\gamma \|\mathbf{x}_1 - \mathbf{x}_2\|^2\right)$
  - sigmoid $k(\mathbf{x}_1, \mathbf{x}_2) = \tanh(\gamma\, \mathbf{x}_1 \cdot \mathbf{x}_2 + c)$

Jeff Howbert Introduction to Machine Learning Winter 2012

**Fig 9: Different Kernels in**
SVM

## III. EXISTING SYSTEM ARCHITECTURE

The approach utilized by [11] with the end goal of highlight extraction was by utilizing Classification And Regression Trees Algorithm. Truck is a paired recursive apportioning process that partitions the scalar qualities and procedures the progressing and discrete attributes. In addition, CART is non-parametric and does not involve a pre-selection of variables because this choice tree can automatically select features [11] . Choice trees are charts that endeavor to show the scope of potential results and ensuing choices settled on after an underlying choice.Decision trees are simple to use compared to other decision-making models, but it is complicated and time-consuming to prepare decision trees, particularly big trees with many branches.

## IV. PROPOSED SYSTEM ARCHITECTURE

A Support Vector Machine (SVM) is authoritatively portrayed by a different hyperplane as an unfair classifier. At the end of the day, given the checked preparing information (regulated learning), a perfect hyperplane is created by the calculation that classifies crisp cases. It divides the two classesrelatively. Every point left of the row falls into the class of the black circle and on the right into the class of the purple square. Separation of courses. This is what SVM is doing .CatBoost is a gradient-boosting algorithm for decision trees. It is created by Yandex scientists and technicians and is used at Yandex and in other businesses, including CERN, Cloudflare, Careem taxi, for search, recommendation systems, personal assistant, self-driving cars, weather prediction and many other activities. It is in open-source and can be used by anyone. CatBoost has several methods to handle categorical characteristics We don't need any unique therapy for one-hot encoded characteristics— a divided search strategy based on histogram can be readily implemented in such cases. During the pre-processing phase, statistical calculation for single categorical characteristics could also be performed.

## V. RESULT

The proposed system is an approach to improve the accuracy of the Earthquake Prediction System as compared to the Algorithms used in previous Research. The methods used here are used to process Time Series Earthquake data recorded as Seismic Signals to predict the time of Failure or the Time at which an Earthquake occurs. To the Seismic Signals we apply Feature Engineering and get Features like the Mean, Standard Deviation, Skew and Quantilewhich can be used to improve the accuracy or reduce the Error of the Solution. Feature Engineering here plays a very important role as a wide range of features allows our Algorithms to learn faster and more efficiently, providing us with more accurate results. CatBoost is a gradient boosting method which is basically an ensemble of decision Trees which is used as a Training Algorithm in the Research. Support Vector Regression is another excellent Algorithm used to Train on the newly Generated Data. Both Algorithms yield us a Mean Absolute Error of 1.74357544138688 and a Cross Validation Score of -2.1651 respectively. This paper therefore, presents a new way to predict Earthquake time. The Research provided in this paper might be further used in systems which might help civilians and Rescue teams to take action before hand. The concluded of our processed model is to improve the accuracy of an Earthquake Time Prediction System and to provide better features for other Systems working in the same Field.

## REFERENCES

1. V.M> Zobin, Introduction to volcanic seismology. Elsevier,2011.
2. W.H. Lee, P. Jennings ,C. Kissinger and H.Kanamori. International handbook of earthquake and engineering seismology. Academic Press,@002,Vol.81.
3. Y. Peng, R. Lahusen, B. Shriazi and W.Song, “Design of smart sensing component for volcano monitoring.” in IET 4th International Conference on Intelligent Environments,July 2008,pp.1-7.
4. R.P.Duin,M.Orozoo-Alzate and J.M. Londono-Bornilla, “Classi-~ fication of volcano events observed by multiple seismic stations,” in 2-0th International Conference on Pattern Recgnition(ICPR),2010,pp.1052-1055.
5. R. Lara-ueva, P. Bernal, M.G. Saltos,D. Ben’itez and J.L Roiolvarez. “Time and frequency feature selection for seismic events from cotopaxi volcano.” in Asia-Pacific Conference on Computer Aided System Engineering(APCASE).2014,2015,pp.129-134.
6. D. Cardenas-Pe’na,M. Orozoo-Alzate, and G. Castellanos-Dominguez,”Selection of time-variant features for earthquake classification at the nevado-del-ruiz volcano, “Computers & Geosciences,vol. 51,pp.293-304,2013
7. B.J. Restrepo, M. Alvarez , and R.Henao, “Comparison between gen- ‘ erative and discriminative approaches for seismic events classification-Comparaioin entre enfoques generativos y discrimativos para la’ classification de eventos s ‘’ ismicos, “Scientia et Technica,vol.1,no.37,2007.

8. R.A. Lara-cueva,D.S Bentez E.V. Carrera,M.Ruiz, and J.L> Roig-Ivarez,"Automatic recognition of long period events from volcano earthquakes at cotopaxi volcano, "IEEE Transactions on Geoscience and Remote Sensing,vol.54,no.9.pp.5247-5257,2016.
9. G.Curilem,J. Vergara,G. Fuentealba, G.Agnua and M.Chag "on ' "Classification of seismic signals at Villarrica volcano(Chile) using neural networks and genetic algorithms, "Journal of Volcanology and Geothermal research,vol.180,no.1,pp.1-8,2009.
10. R.Lara-cueva,D.Ben'itez,E.Carrera,M.Ruiz,an d J.Roig-Alvaraz,' "Feature selection of seismic waveforms for long period event detection at cotopaxi volcano, "Journal of Volocanlogy and Geothernal research, vol.316,pp.34-49,2016.
11. Precursory Pattern based Feature Extraction Techniques for Earthquake Prediction. Lei Zhang, LangchunSi, Haipeng yang, Yuanzhi Hu, Jianfeng Qiu.

## AUTHORS PROFILE

**Sahaya Sakila,** Assistant Professor, Computer Science And Engineering Department, in SRM Institute of Science And Technology, Ramapuram, Chennai

**Sanyam Garg, ,**Pre-Final Year Student of B.Tech, Computer Science And Engineering in SRM Institute of Science And Technology, Ramapuram, Chennai
.

**Tanay Yeole**, Pre-Final Year Student of B.Tech, Computer Science And Engineering in SRM Institute of Science And Technology, Ramapuram, Chennai

**Hrithik Yadav,** Pre-Final Year Student of B.Tech, Computer Science And Engineering in SRM Institute of Science And Technology, Ramapuram, Chennai