International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 05 | May 2021 www.irjet.net p-ISSN: 2395-0072

Earthquake Prediction using Machine Learning

Dr. S. Anbu Kumar[1], Abhay Kumar[2], Aditya Dhanraj[3], Ashish Thakur[4]

[1-4]DEPARTMENT OF CIVIL ENGINEERING, DELHI TECHNOLOGICAL UNIVERSITY, (Formerly Delhi College of Engineering), Bawana Road, New Delhi-110042

---***---

ABSTRACT: During this study, earthquake prediction was performed, by training different Machine Learning models on seismic and acoustic data collected from a laboratory micro-earthquake simulation. Prediction has been made by extracting 40 statistical features, such as no. of peaks, time to failure etc. from the 'single-feature' acoustic data, which was basically in the form of a time series. During this research, six machine learning techniques including Linear Regression, Support Vector Machine, Random Forest Regression, Case Based Reasoning, XGBoost and Light Gradient Boosting Mechanism are separately applied and accuracies in the training and testing datasets were compared to pick out the best model. Furthermore, the evaluation of accuracy is another step taken into account for analysing the result. The above methods for predicting earthquake magnitude yield significant and encouraging results, signalling advancement toward the ultimate robust prediction process.

Keywords: Earthquake Prediction, Machine Learning, Regressors.

CHAPTER 1

INTRODUCTION

Natural disasters result in a large number of deaths, property loss, damages and injuries. Individuals cannot avoid them, but early prediction and appropriate protective precautions can minimize human life casualties and save a large number of valuable items. Earthquake is one amongst the main such disaster. Presently, we don't have any specific technique that can be used for predicting earthquake, unlike other disaster, that makes it much more devastating. Some researchers believe that earthquakes can't be anticipated, whereas others believe they are a predictable occurrence. According to them, many procedures for earthquake prediction are often used, including the study of quick visual phenomena such as changes in electric field, magnetic field, total electron content of the ionosphere, change in animal behaviour and historic earthquake records, all of which are well kept in the form of collection. A model capable of predicting earthquakes must be able to predict the accurate location, magnitude spectrum and precise occurrence time and chances of occurrence. Until now, there has not been a comprehensive way to predict earthquake. Indeed, an earthquake prediction mechanism that provides precise prediction is urgently needed. A signal created by such a device could allow authorities to deploy resources, and shutdown devices which will cause major damage like atomic power plants & power grid so that deaths and damages can be avoid. The input parameter for this earthquake prediction study were derived from a laboratory micro earthquake simulation. These types of steaky distributions show the frequency of laboratory micro earthquake simulation events as function of magnitudes. These function and distinct parameters are used to figure out the fundamental relationship between geophysical activity of seismic tranquilly and major earthquake frequency. Irrespective of degree of the nonlinearity among them, the relationship between seismic activity and geophysical data must be modelled. Seismic contemplation is a break in the natural release of seismic energy obtain from fracture region. These concentration of seismic energy inside the faults region may result in earthquake. Amount of seismic energy stored can be used to estimate the magnitude of next coming earthquakes. Similarly, major earthquake frequency is taken into account as a precursor of a major earthquake. Major earthquakes are the sequence of earthquakes, which has magnitude significantly higher frequency than the previous seismic activity. Machine Learning (ML) is employed in fields for the purpose of prediction and categorization. The main idea of this project is to depict the time that we have before laboratory earthquake occurred from real time seismic data. These laboratory seismic data are used for the purpose of input to the various Machine Learning approaches. These include Random Forest Regression, Linear Regression, Light Gradient Boosting Mechanism, Support Vector Machine, Case Based Reasoning and XGBoost ensemble of decision trees to predict earthquake. During this paper we have extract the data from all the above mention techniques and we also compared these techniques so that we come to a conclusion that which technique is best for predicting earthquake.

CHAPTER 2

LITERATURE REVIEW

Earthquake activity is presumed as a spontaneous phenomenon that can damage huge number of lives and properties, and currently there is no any model exists that can predict the exact position, magnitude, frequency and time of an earthquake. Researchers have conducted several experiments on earthquake events and forecasts, leading to a variety of findings based on the factors considered. The well-known Gutenberg and Richter statistical model found a correlation between the magnitude of earthquake and frequency of earthquake. For structural design, this earthquake probability distribution model was used. In supervision of the California Geological Survey, Petersen conducted research and suggested a model that is time-independent. This time independent model demonstrating that chances of occurrence of earthquake follow the Poisson's distribution model. Shen suggested a probabilistic earthquake forecasting model based on the strain studied between the behaviour of tectonic plates. Based on this model, higher measured strain results in a higher risk of earthquake. Ebel provided a long-term prediction model that allowed for the extrapolation of previous earthquakes with magnitudes greater than and up to 5.2 in order to forecast possible seismic events.

There are various methods for predicting earthquakes using Artificial Neural Networks and seismic precursors are discussed in the literature. Negarestani used a Back Propagation Neural Network to identify irrational behaviour in concentration of radon due to occurrence of earthquake. The presence of radon gas in soil is constantly measured and researcher have founded that it varies constantly due to changes in environment. The concentration of soil radon also rises due to seismic activity. This radon can be differentiated from natural variations caused by the environment through neural networks. Since splitting the entire globe in four quadrants, the system devices establish logic and correlation principles based on the historical record of earthquakes. The expert method will forecast earthquakes in each quadrant of the world for a period of 24 hours.

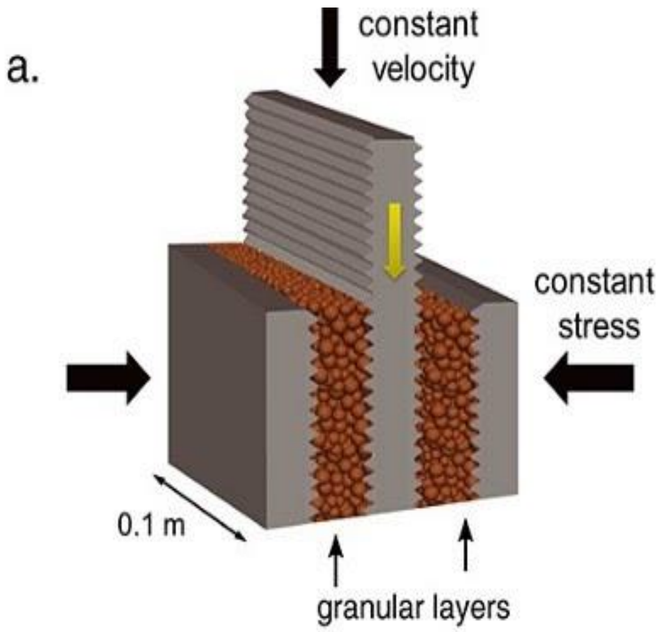
Panakkat and Adeli presented an enthralling approach to earthquake prediction based on mathematically determined seismic indicators derived from the spatial variation of historical seismic events for Southern California. The algorithm makes monthly predictions, and the parameters are modelled using various Artificial Neural Networks. The estimation of all those parameters required to make sufficient earthquake database. For this limited number of times, the events were executed to measure the parameters of seismic event before taking the month into account. After this study, Adeli and Panakkat used exactly same parameters of seismic in collaboration with Probabilistic Neural Network to forecast earthquakes.

Morales-Esteban and Reyes suggested separate seismic criteria for earthquake prediction using mathematical calculations in Chile and Iberia for a time interval of 8–9 days, respectively. For modelling the relationship between earthquake events and parameters, these parameters are determined using Bath's law and Omori's law. Zamani proposes using a combination of neural networks and mathematical logic to forecast earthquakes in Iran. For a selected group of seismicity indices, this study includes information normalization and corresponding feature extraction accompanied by principal component analysis. Mirrashid provides another design for earthquake prediction in Iran, which incorporates symbolic logic, fuzzy C-means, subtractive clustering, and grid partitioning. Through this model, we try to predict earthquakes by training various Machine Learning models on seismic and acoustic data from a laboratory micro earthquake simulation.

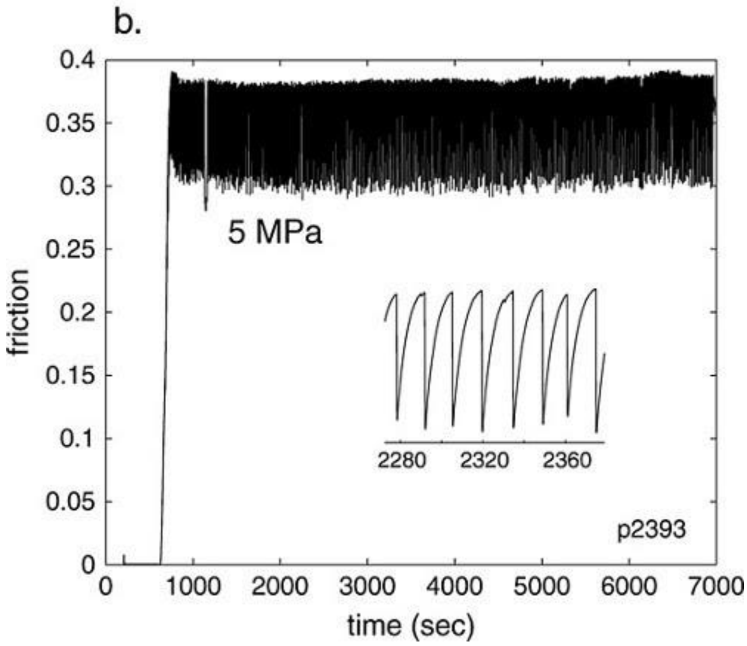
CHAPTER 3

SIMULATED EARTHQUAKE ENVIRONMENT

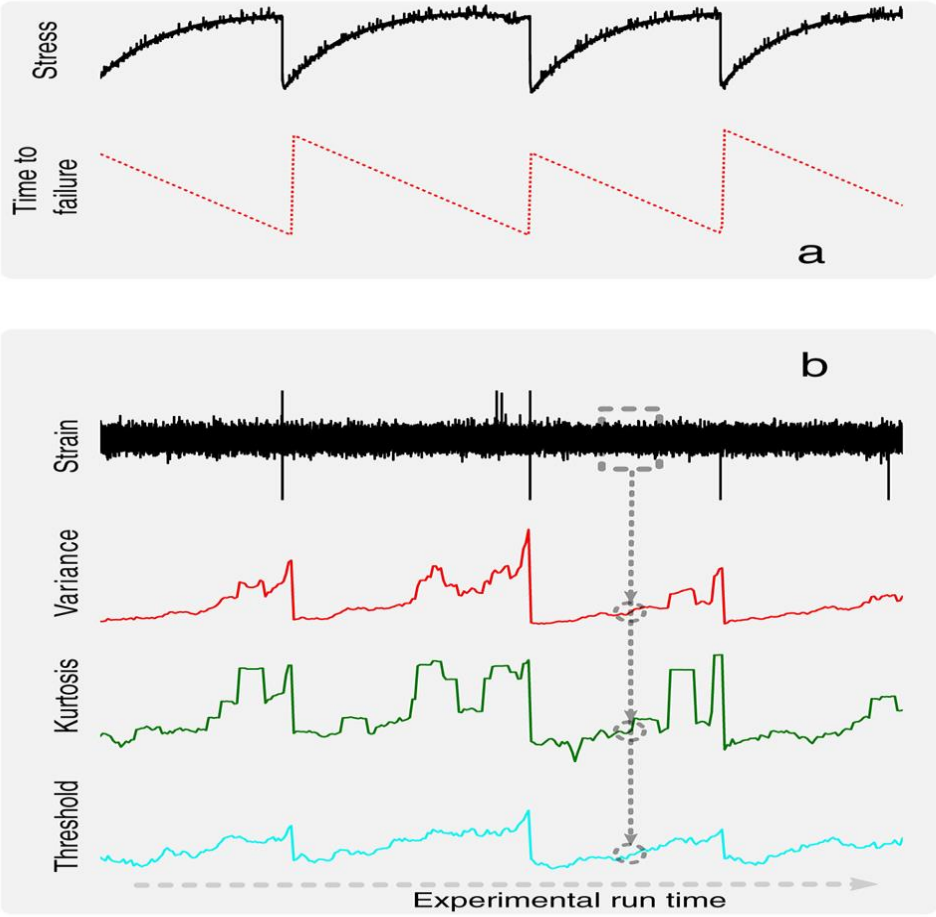
The data that we are using came from an experiment that was conducted on rock during a very double direct shear geometry which was subjected to bi-axial packing, in classic laboratory earthquake model. Two fault gouge layers were sheared simultaneously while plagued to a relentless normal load and a mentioned shear velocity.



The laboratory faults fail in repetitive cycles of stick and slip that is meant to mimic the cycle of loading and failure on tectonic faults. While the experiment is considerably simplified than a fault on Earth, it shares certain physical characteristics, whose similarity, just cannot be ignored.



When we take small section of repetitive cycle of stick and zoomed it, we got the variance of stress versus time. As shown below:



In case of quasi-periodic laboratory seismic cycles, the prediction of laboratory earthquake from continuous seismic data is possible.

CHAPTER 4

DATA SET

The dimension of the info is sort of large, in way over 600 million rows of information. The two columns within the train dataset have the subsequent meaning:

Acoustic data: is that the acoustic signal measured within the laboratory experiment;

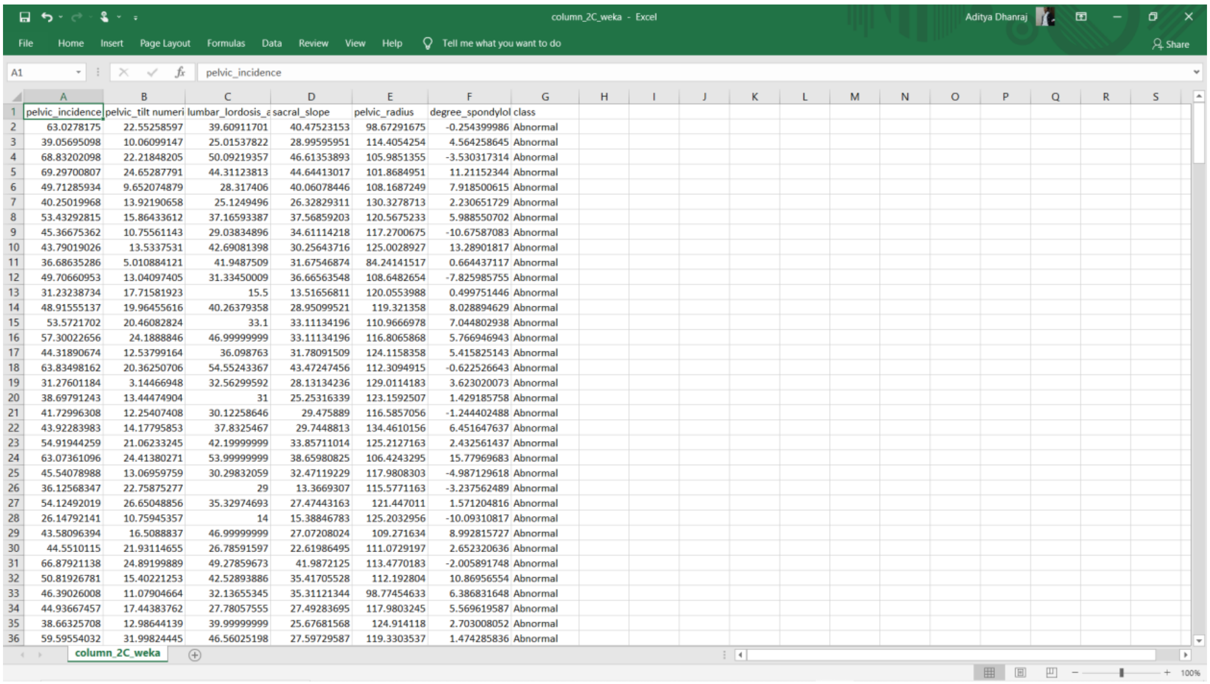
Time to failure: this provides the time until a failure will occur.

We have plotted 1% of the info. For this we are going to sample every 100 points of knowledge.

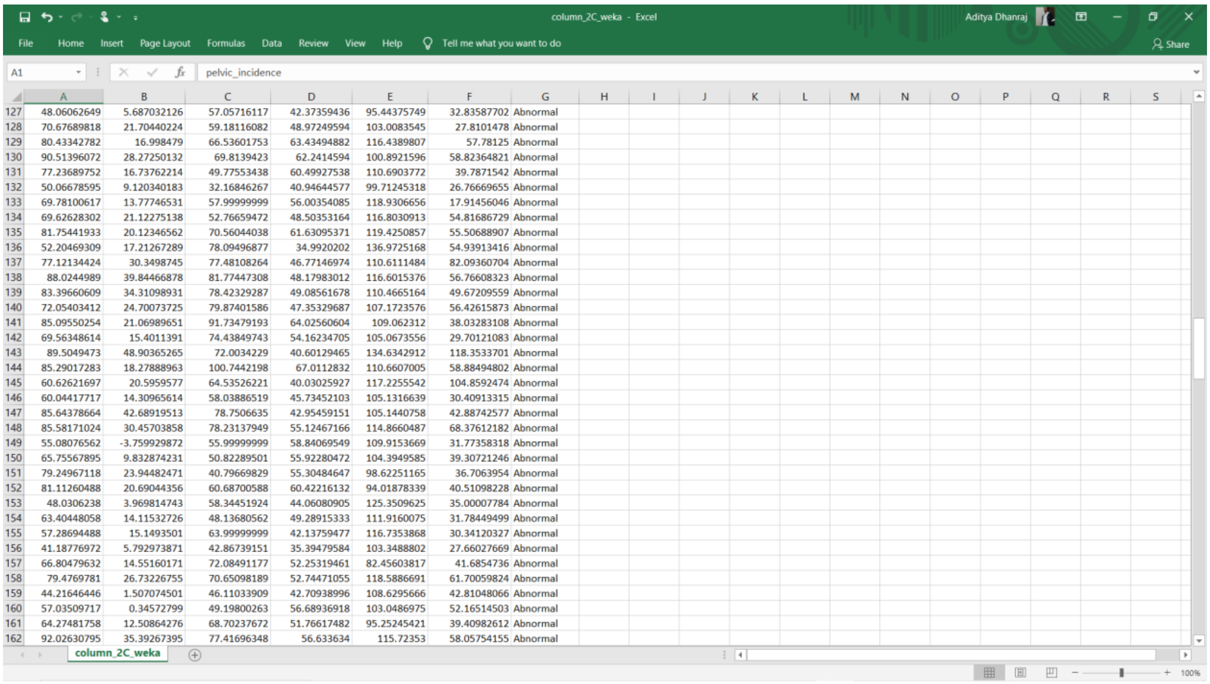
index	acoustic_data	time_to_failure
0	12	1.4691
1	6	1.4691
2	8	1.4691
3	5	1.4691
4	8	1.4691

Fig: First 5 observation of the Dataset

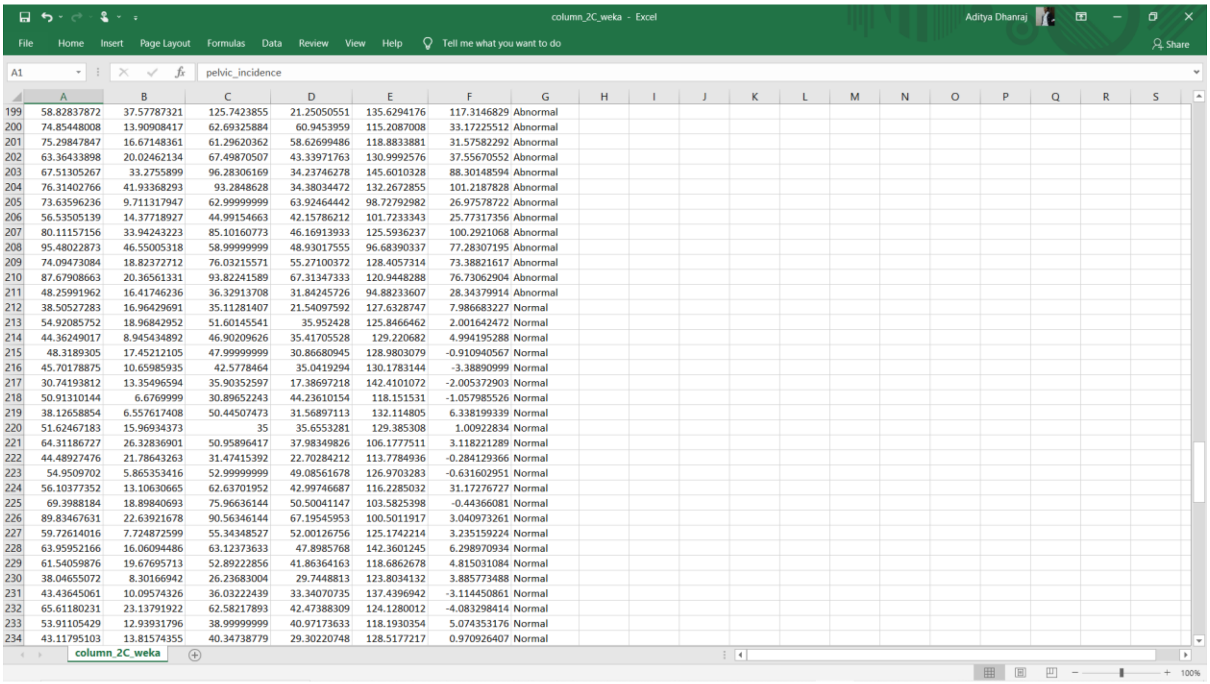
Some sample of data set are shown below:



	A	B	C	D	E	F	G
1	pelvic_incidence	pelvic_tilt numeri	lumbar_lordosis_a	sacral_slope	pelvic_radius	degree_spondylol	class
2	63.0278175	22.55258597	39.60911701	40.47523153	98.67291675	-0.254399986	Abnormal
3	39.05695098	10.06099147	25.01537822	28.99595951	114.4054254	4.564258645	Abnormal
4	68.83202098	22.21848205	50.09219357	46.61353893	105.9851355	-3.530317314	Abnormal
5	69.29700807	24.65287791	44.31123813	44.64413017	101.8684951	11.21152344	Abnormal
6	49.71285934	9.652074879	28.317406	40.06078446	108.1687249	7.918500615	Abnormal
7	40.25019968	13.92190658	25.1249496	26.32829311	130.3278713	2.230651729	Abnormal
8	53.43292815	15.86433612	37.16593387	37.56859203	120.5675233	5.988550702	Abnormal
9	45.36675362	10.75561143	29.03834896	34.61114218	117.2700675	-10.67587083	Abnormal
10	43.79019026	13.5337531	42.69081398	30.25643716	125.0028927	13.28901817	Abnormal
11	36.68635286	5.010884121	41.9487509	31.67546874	84.24141517	0.664437117	Abnormal
12	49.70660953	13.04097405	31.33450009	36.66563548	108.6482654	-7.825985755	Abnormal
13	31.23238734	17.71581923	15.5	13.51656811	120.0553988	0.499751446	Abnormal
14	48.91555137	19.96455616	40.26379358	28.95099521	119.321358	8.028894629	Abnormal
15	53.5721702	20.46082824	33.1	33.11134196	110.9666978	7.044802938	Abnormal
16	57.30022656	24.1888846	46.99999999	33.11134196	116.8065868	5.766946943	Abnormal
17	44.31890674	12.53799164	36.098763	31.78091509	124.1158358	5.415825143	Abnormal
18	63.83498162	20.36250706	54.55243367	43.47247456	112.3094915	-0.622526643	Abnormal
19	31.27601184	3.14466948	32.56299592	28.13134236	129.0114183	3.623020073	Abnormal
20	38.69791243	13.44474904	31	25.25316339	123.1592507	1.429185758	Abnormal
21	41.72996308	12.25407408	30.12258646	29.475889	116.5857056	-1.244402488	Abnormal
22	43.92283983	14.17795853	37.8325467	29.7448813	134.4610156	6.451647637	Abnormal
23	54.91944259	21.06233245	42.19999999	33.85711014	125.2127163	2.432561437	Abnormal
24	63.07361096	24.41380271	53.99999999	38.65980825	106.4243295	15.77969683	Abnormal
25	45.54078988	13.06959759	30.29832059	32.47119229	117.9808303	-4.987129618	Abnormal
26	36.12568347	22.75875277	29	13.3669307	115.5771163	-3.237562489	Abnormal
27	54.12492019	26.65048856	35.32974693	27.47443163	121.447011	1.571204816	Abnormal
28	26.14792141	10.75945357	14	15.38846783	125.2032956	-10.09310817	Abnormal
29	43.58096394	16.5088837	46.99999999	27.07208024	109.271634	8.992815727	Abnormal
30	44.5510115	21.93114655	26.78591597	22.61986495	111.0729197	2.652320636	Abnormal
31	66.87921138	24.89199889	49.27859673	41.9872125	113.4770183	-2.005891748	Abnormal
32	50.81926781	15.40221253	42.52893886	35.41705528	112.192804	10.86956554	Abnormal
33	46.39026008	11.07904664	32.13655345	35.31121344	98.77454633	6.386831648	Abnormal
34	44.93667457	17.44383762	27.78057555	27.49283695	117.9803245	5.569619587	Abnormal
35	38.66325708	12.98644139	39.99999999	25.67681568	124.914118	2.703008052	Abnormal
36	59.59554032	31.99824445	46.56025198	27.59729587	119.3303537	1.474285836	Abnormal

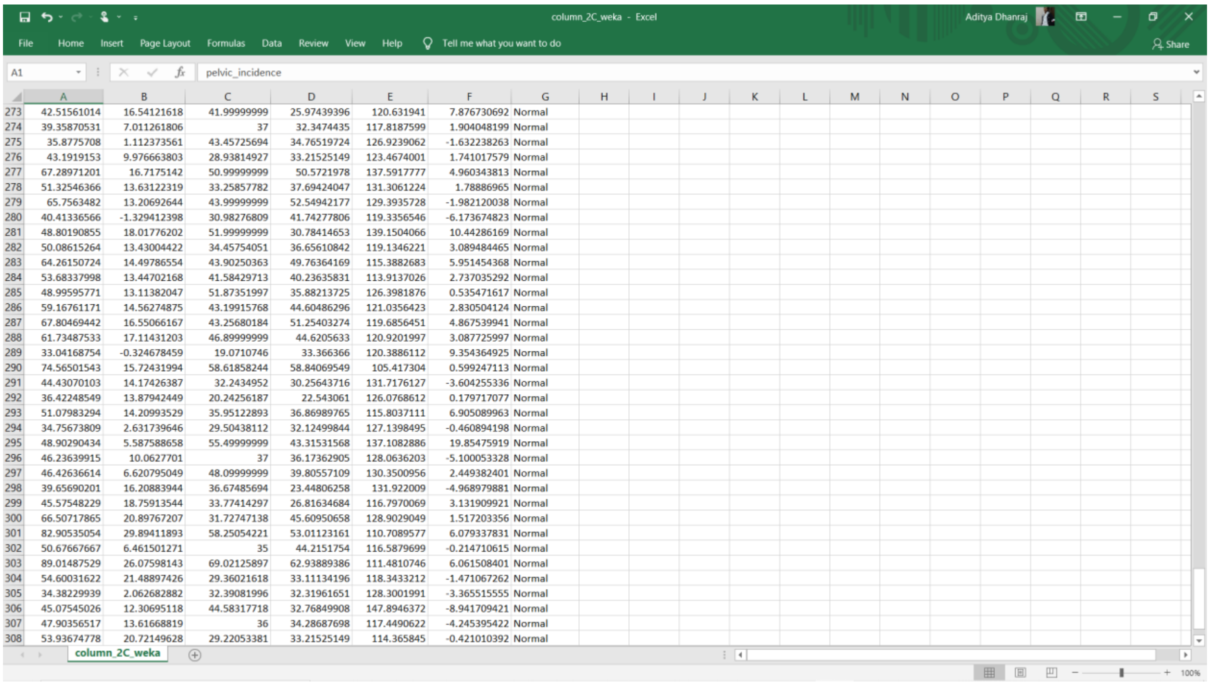


	A	B	C	D	E	F	G
127	48.06062649	5.687032126	57.05716117	42.37359436	95.44375749	32.83587702	Abnormal
128	70.67689818	21.70440224	59.18116082	48.97249594	103.0083545	27.8101478	Abnormal
129	80.43342782	16.998479	66.53601753	63.43494882	116.4389807	57.78125	Abnormal
130	90.51396072	28.27250132	69.8139423	62.2414594	100.8921596	58.82364821	Abnormal
131	77.23689752	16.73762214	49.77553438	60.49927538	110.6903772	39.7871542	Abnormal
132	50.06678595	9.120340183	32.16846267	40.94644577	99.71245318	26.76669655	Abnormal
133	69.78100617	13.77746531	57.99999999	56.00354085	118.9306656	17.91456046	Abnormal
134	69.62628302	21.12275138	52.76659472	48.50353164	116.8030913	54.81686729	Abnormal
135	81.75441933	20.12346562	70.56044038	61.63095371	119.4250857	55.50688907	Abnormal
136	52.20469309	17.21267289	78.09496877	34.9920202	136.9725168	54.93913416	Abnormal
137	77.12134424	30.3498745	77.48108264	46.77146974	110.6111484	82.09360704	Abnormal
138	88.0244989	39.84466878	81.77447308	48.17983012	116.6015376	56.76608323	Abnormal
139	83.39660609	34.31098931	78.42329287	49.08561678	110.4665164	49.67209559	Abnormal
140	72.05403412	24.70073725	79.87401586	47.35329687	107.1723576	56.42615873	Abnormal
141	85.09550254	21.06989651	91.73479193	64.02560604	109.062312	38.03283108	Abnormal
142	69.56348614	15.4011391	74.43849743	54.16234705	105.0673556	29.70121083	Abnormal
143	89.5049473	48.90365265	72.0034229	40.60129465	134.6342912	118.3533701	Abnormal
144	85.29017283	18.27888963	100.7442198	67.0112832	110.6607005	58.88494802	Abnormal
145	60.62621697	20.5959577	64.53526221	40.03025927	117.2255542	104.8592474	Abnormal
146	60.04417717	14.30965614	58.03886519	45.73452103	105.1316639	30.40913315	Abnormal
147	85.64378664	42.68919513	78.7506635	42.95459151	105.1440758	42.88742577	Abnormal
148	85.58171024	30.45703858	78.23137949	55.12467166	114.8660487	68.37612182	Abnormal
149	55.08076562	-3.759929872	55.99999999	58.84069549	109.9153669	31.77358318	Abnormal
150	65.75567895	9.832874231	50.82289501	55.92280472	104.3949585	39.30721246	Abnormal
151	79.24967118	23.94482471	40.79669829	55.30484647	98.62251165	36.7063954	Abnormal
152	81.11260488	20.69044356	60.68700588	60.42216132	94.01878339	40.51098228	Abnormal
153	48.0306238	3.969814743	58.34451924	44.06080905	125.3509625	35.00007784	Abnormal
154	63.40448058	14.11532726	48.13680562	49.28915333	111.9160075	31.78449499	Abnormal
155	57.28694488	15.1493501	63.99999999	42.13759477	116.7353868	30.34120327	Abnormal
156	41.18776972	5.792973871	42.86739151	35.39479584	103.3488802	27.66027669	Abnormal
157	66.80479632	14.55160171	72.08491177	52.25319461	82.45603817	41.6854736	Abnormal
158	79.4769781	26.73226755	70.65098189	52.74471055	118.5886691	61.70059824	Abnormal
159	44.21646446	1.507074501	46.11033909	42.70938996	108.6295666	42.81048066	Abnormal
160	57.03509717	0.34572799	49.19800263	56.68936918	103.0486975	52.16514503	Abnormal
161	64.27481758	12.50864276	68.70237672	51.76617482	95.25245421	39.40982612	Abnormal
162	92.02630795	35.39267395	77.41696348	56.633634	115.72353	58.05754155	Abnormal



	A	B	C	D	E	F	G
199	58.82837872	37.57787321	125.7423855	21.25050551	135.6294176	117.3146829	Abnormal
200	74.85448008	13.90908417	62.69325884	60.9453959	115.2087008	33.17225512	Abnormal
201	75.29847847	16.67148361	61.29620362	58.62699486	118.8833881	31.57582292	Abnormal
202	63.36433898	20.02462134	67.49870507	43.33971763	130.9992576	37.55670552	Abnormal
203	67.51305267	33.2755899	96.28306169	34.23746278	145.6010328	88.30148594	Abnormal
204	76.31402766	41.93368293	93.2848628	34.38034472	132.2672855	101.2187828	Abnormal
205	73.63596236	9.711317947	62.99999999	63.92464442	98.72792982	26.97578722	Abnormal
206	56.53505139	14.37718927	44.99154663	42.15786212	101.7233343	25.77317356	Abnormal
207	80.11157156	33.94243223	85.10160773	46.16913933	125.5936237	100.2921068	Abnormal
208	95.48022873	46.55005318	58.99999999	48.93017555	96.68390337	77.28307195	Abnormal
209	74.09473084	18.82372712	76.03215571	55.27100372	128.4057314	73.38821617	Abnormal
210	87.67908663	20.36561331	93.82241589	67.31347333	120.9448288	76.73062904	Abnormal
211	48.25991962	16.41746236	36.32913708	31.84245726	94.88233607	28.34379914	Abnormal
212	38.50527283	16.96429691	35.11281407	21.54097592	127.6328747	7.986683227	Normal
213	54.92085752	18.96842952	51.60145541	35.952428	125.8466462	2.001642472	Normal
214	44.36249017	8.945434892	46.90209626	35.41705528	129.220682	4.994195288	Normal
215	48.3189305	17.45212105	47.99999999	30.86680945	128.9803079	-0.910940567	Normal
216	45.70178875	10.65985935	42.5778464	35.0419294	130.1783144	-3.38890999	Normal
217	30.74193812	13.35496594	35.90352597	17.38697218	142.4101072	-2.005372903	Normal
218	50.91310144	6.6769999	30.89652243	44.23610154	118.151531	-1.057985526	Normal
219	38.12658854	6.557617408	50.44507473	31.56897113	132.114805	6.338199339	Normal
220	51.62467183	15.96934373	35	35.6553281	129.385308	1.00922834	Normal
221	64.31186727	26.32836901	50.95896417	37.98349826	106.1777511	3.118221289	Normal
222	44.48927476	21.78643263	31.47415392	22.70284212	113.7784936	-0.284129366	Normal
223	54.9509702	5.865353416	52.99999999	49.08561678	126.9703283	-0.631602951	Normal
224	56.10377352	13.10630665	62.63701952	42.99746687	116.2285032	31.17276727	Normal
225	69.3988184	18.89840693	75.96636144	50.50041147	103.5825398	-0.44366081	Normal
226	89.83467631	22.63921678	90.56346144	67.19545953	100.5011917	3.040973261	Normal
227	59.72614016	7.724872599	55.34348527	52.00126756	125.1742214	3.235159224	Normal
228	63.95952166	16.06094486	63.12373633	47.8985768	142.3601245	6.298970934	Normal
229	61.54059876	19.67695713	52.89222856	41.86364163	118.6862678	4.815031084	Normal
230	38.04655072	8.30166942	26.23683004	29.7448813	123.8034132	3.885773488	Normal
231	43.43645061	10.09574326	36.03222439	33.34070735	137.4396942	-3.114450861	Normal
232	65.61180231	23.13791922	62.58217893	42.47388309	124.1280012	-4.083298414	Normal
233	53.91105429	12.93931796	38.99999999	40.97173633	118.1930354	5.074353176	Normal
234	43.11795103	13.81574355	40.34738779	29.30220748	128.5177217	0.970926407	Normal

International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 05 | May 2021 www.irjet.net p-ISSN: 2395-0072



	A	B	C	D	E	F	G
273	42.51561014	16.54121618	41.99999999	25.97439396	120.631941	7.876730692	Normal
274	39.35870531	7.011261806	37	32.3474435	117.8187599	1.904048199	Normal
275	35.8775708	1.112373561	43.45725694	34.76519724	126.9239062	-1.632238263	Normal
276	43.1919153	9.976663803	28.93814927	33.21525149	123.4674001	1.741017579	Normal
277	67.28971201	16.7175142	50.99999999	50.5721978	137.5917777	4.960343813	Normal
278	51.32546366	13.63122319	33.25857782	37.69424047	131.3061224	1.78886965	Normal
279	65.7563482	13.20692644	43.99999999	52.54942177	129.3935728	-1.982120038	Normal
280	40.41336566	-1.329412398	30.98276809	41.74277806	119.3356546	-6.173674823	Normal
281	48.80190855	18.01776202	51.99999999	30.78414653	139.1504066	10.44286169	Normal
282	50.08615264	13.43004422	34.45754051	36.65610842	119.1346221	3.089484465	Normal
283	64.26150724	14.49786554	43.90250363	49.76364169	115.3882683	5.951454368	Normal
284	53.68337998	13.44702168	41.58429713	40.23635831	113.9137026	2.737035292	Normal
285	48.99595771	13.11382047	51.87351997	35.88213725	126.3981876	0.535471617	Normal
286	59.16761171	14.56274875	43.19915768	44.60486296	121.0356423	2.830504124	Normal
287	67.80469442	16.55066167	43.25680184	51.25403274	119.6856451	4.867539941	Normal
288	61.73487533	17.11431203	46.89999999	44.6205633	120.9201997	3.087725997	Normal
289	33.04168754	-0.324678459	19.0710746	33.366366	120.3886112	9.354364925	Normal
290	74.56501543	15.72431994	58.61858244	58.84069549	105.417304	0.599247113	Normal
291	44.43070103	14.17426387	32.2434952	30.25643716	131.7176127	-3.604255336	Normal
292	36.42248549	13.87942449	20.24256187	22.543061	126.0768612	0.179717077	Normal
293	51.07983294	14.20993529	35.95122893	36.86989765	115.8037111	6.905089963	Normal
294	34.75673809	2.631739646	29.50438112	32.12499844	127.1398495	-0.460894198	Normal
295	48.90290434	5.587588658	55.49999999	43.31531568	137.1082886	19.85475919	Normal
296	46.23639915	10.0627701	37	36.17362905	128.0636203	-5.100053328	Normal
297	46.42636614	6.620795049	48.09999999	39.80557109	130.3500956	2.449382401	Normal
298	39.65690201	16.20883944	36.67485694	23.44806258	131.922009	-4.968979881	Normal
299	45.57548229	18.75913544	33.77414297	26.81634684	116.7970069	3.131909921	Normal
300	66.50717865	20.89767207	31.72747138	45.60950658	128.9029049	1.517203356	Normal
301	82.90535054	29.89411893	58.25054221	53.01123161	110.7089577	6.079337831	Normal
302	50.67667667	6.461501271	35	44.2151754	116.5879699	-0.214710615	Normal
303	89.01487529	26.07598143	69.02125897	62.93889386	111.4810746	6.061508401	Normal
304	54.60031622	21.48897426	29.36021618	33.11134196	118.3433212	-1.471067262	Normal
305	34.38229939	2.062682882	32.39081996	32.31961651	128.3001991	-3.365515555	Normal
306	45.07545026	12.30695118	44.58317718	32.76849908	147.8946372	-8.941709421	Normal
307	47.90356517	13.61668819	36	34.28687698	117.4490622	-4.245395422	Normal
308	53.93674778	20.72149628	29.22053381	33.21525149	114.365845	-0.421010392	Normal

CHAPTER 5

EXPLORATORY DATA ANALYSIS

It is impossible to plot graph of every data that we have collected. That's why we have decided to shows only one of the total data. The acoustic data shows very complex oscillations with variable amplitude. On plotting both the data i.e. Time to failure and total Acoustic Data on a single plot, we have,

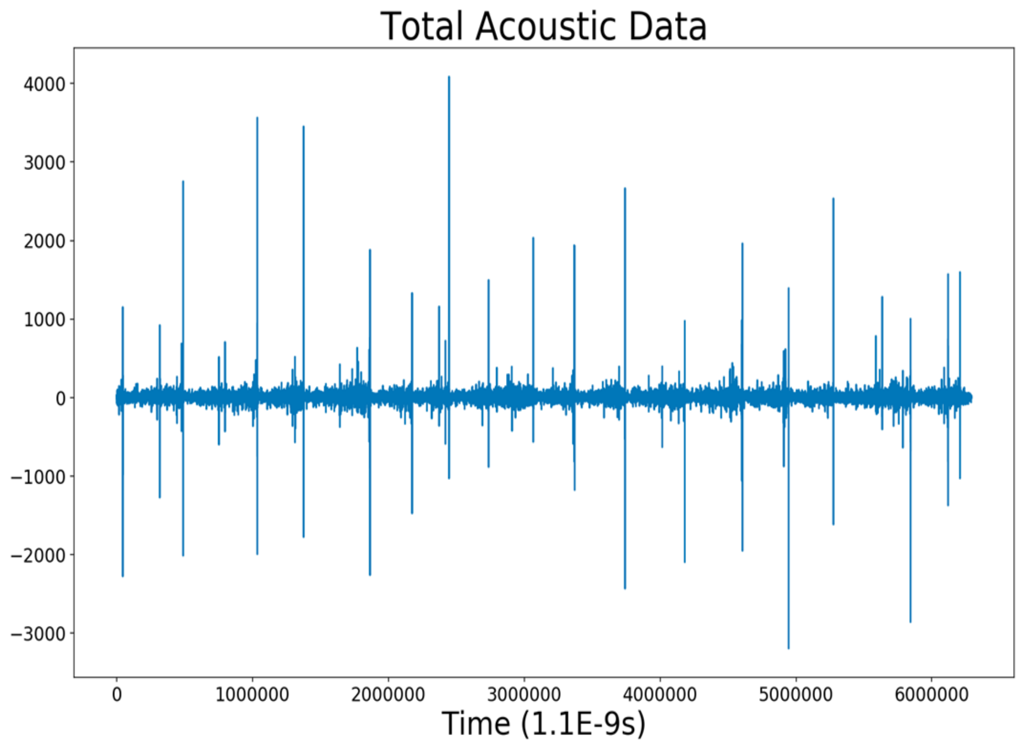
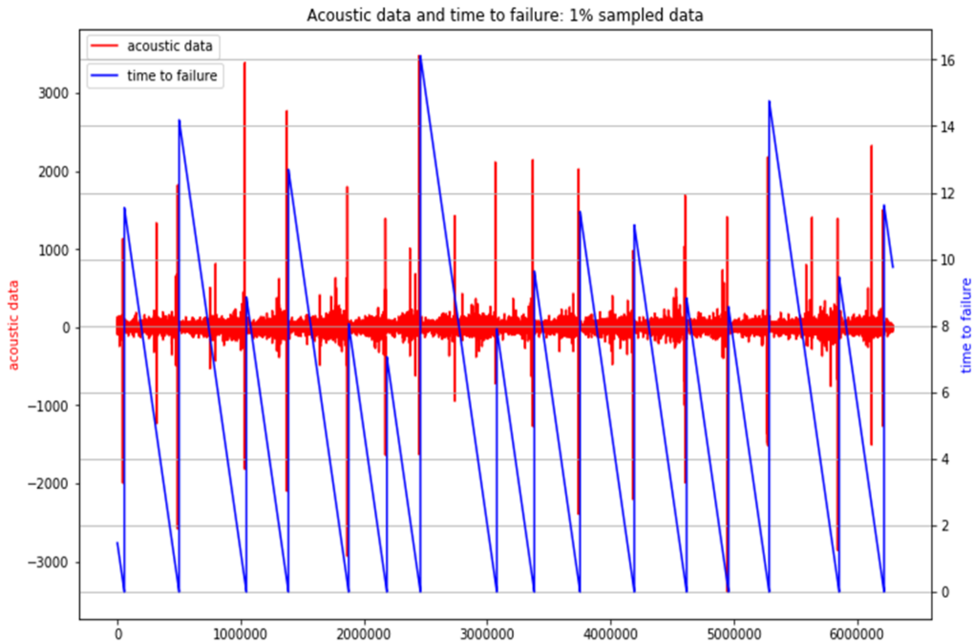


Fig: Total Acoustic Data, plotted against time

Just before each failure there's an amplitude rise in the acoustic data. We also see that numerous amplitudes have been observed in different moments in time (for e.g. about the mid-time between two consecutive failures). We plot similarly the primary 1% i.e. the first 1 % of the data to get a zoomed view.



On this zoomed-in-time plot, we are able to see that really the massive oscillation before the failure isn't quite within the last moment. There are a chain of high frequency oscillations before the big one, and also some low amplitude peaks following it. This is again followed by some minor oscillations before the occurrence of failure. This pattern is observed almost throughout the data and guides us to our hypothesis, and we performed feature engineering and model to test the same.

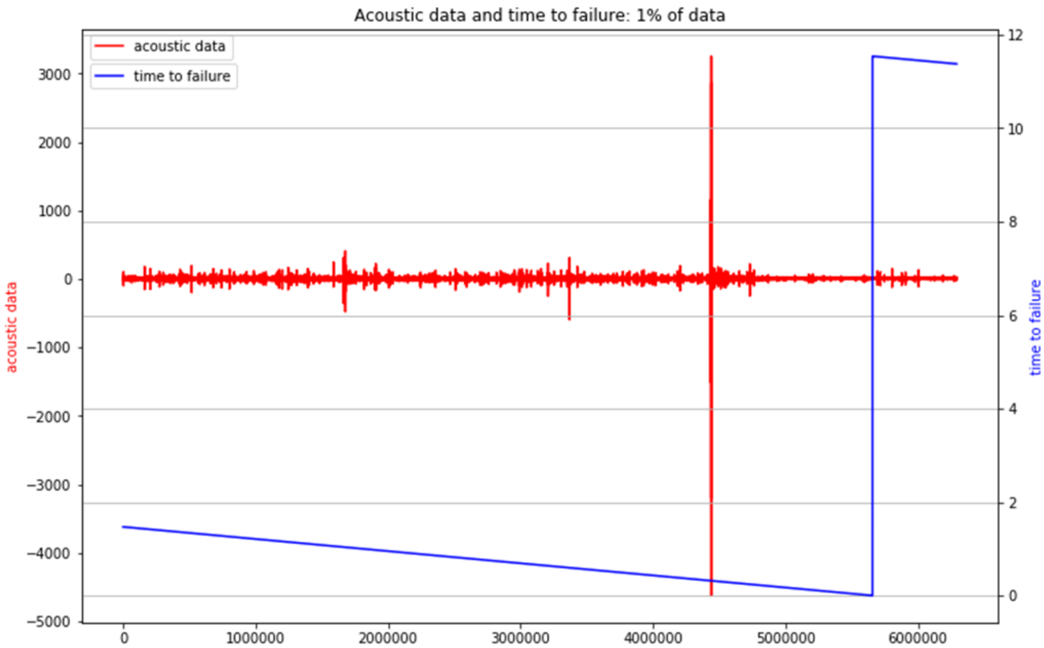


Fig: First 1% of Acoustic Data (Red) and Time to failure (Blue) against time.

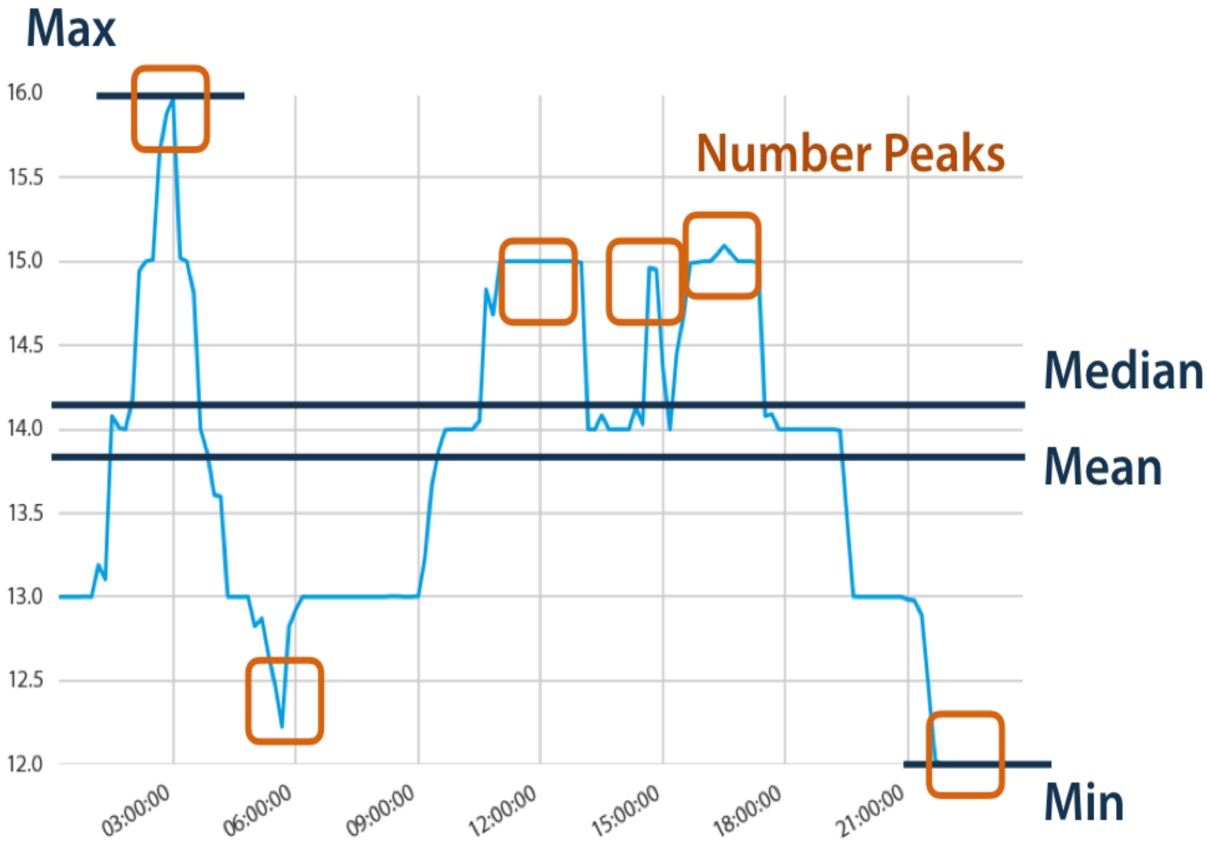
CHAPTER 6

FEATURES ENGINEERING

Test segment has more than set of 1,50,000 data. For our convenience we would take one out of 100th observation in the model.

Now, after pre-processing the data, we are encountered with a new problem, that how do we going to solve this as a regression problem of this acoustic with a single feature. This type of problems is very popular among the data scientists whose attempt is to make forecasts or try to detect signals in time sequence.

For this we deployed some statistical methodologies to extract some basic aggregate features such as max, min, median, standard deviation, segment's mean, IQR etc, especially in time series analysis.

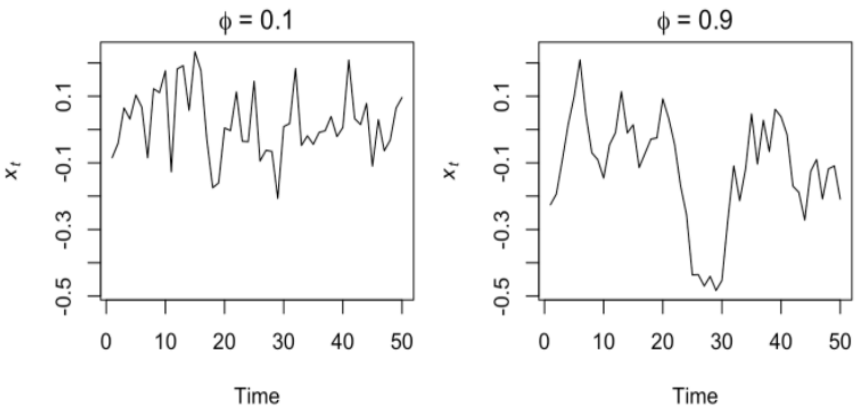


If we extracted these features from time series data, the problem converted to merely a machine learning problem. We extracted 700 such features from 200 segments of 150000 observations in total for our model. Some of them are:

6.1 Auto-Regressive Model Coefficient

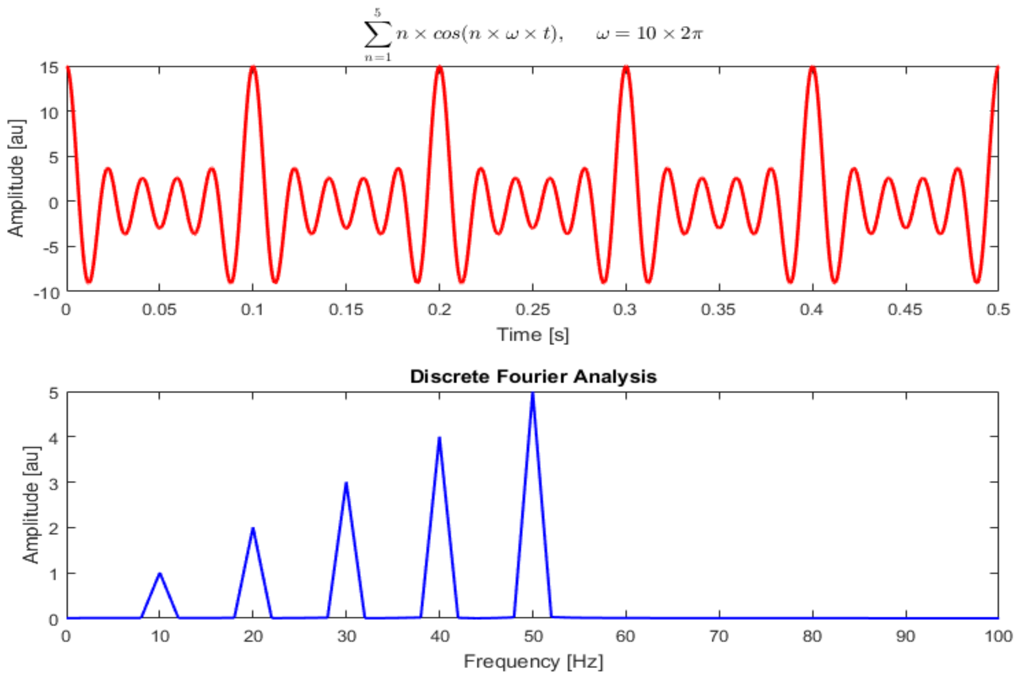
Through this model, we transform a sequence of time in a regression problem in which previous values are used in the form of time sequences. This time sequences are features and coefficients of regression model eventually proves to be a crucial features for my Light Gradient Boosting Mechanism (LGBM), which shows that time sequences acoustic data actually have an element of lag.

```r
## setup plot region
par(mfrow = c(1, 2))
## get y-limits for common plots
ylm <- c(min(AR1_sm, AR1_lg), max(AR1_sm, AR1_lg))
## plot the ts
plot.ts(AR1_sm, ylim = ylm, ylab = expression(italic(x)[italic(t)]),
    main = expression(paste(phi, " = 0.1")))
plot.ts(AR1_lg, ylim = ylm, ylab = expression(italic(x)[italic(t)]),
    main = expression(paste(phi, " = 0.9")))
```



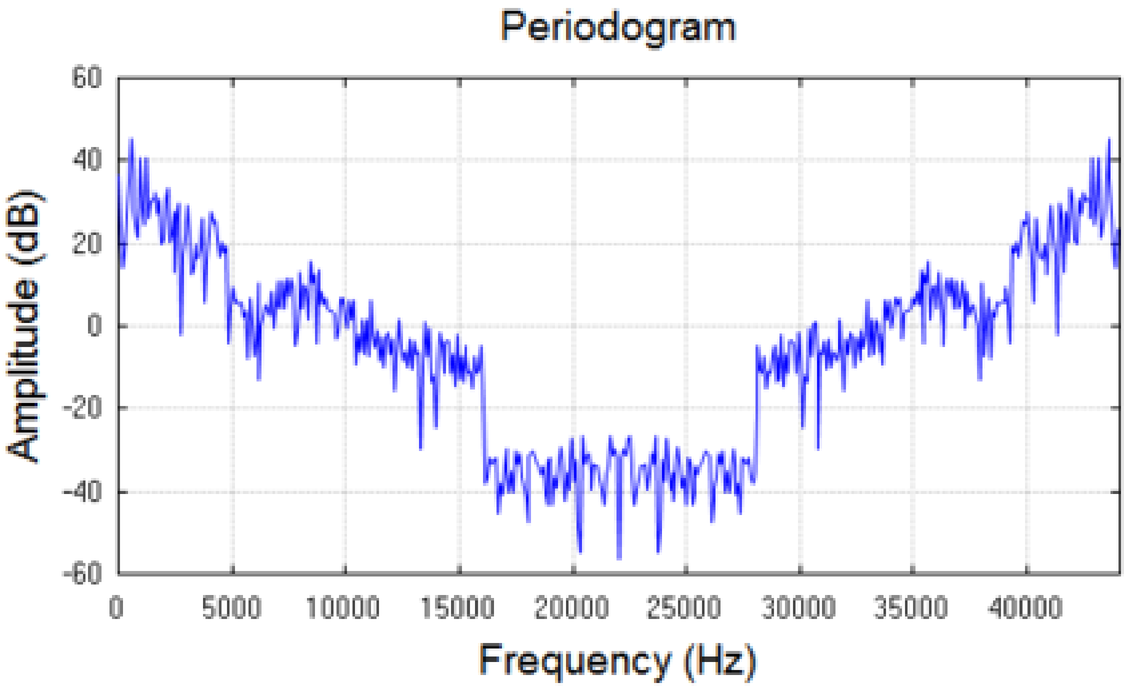
6.2 Fast Fourier Transformation Variance

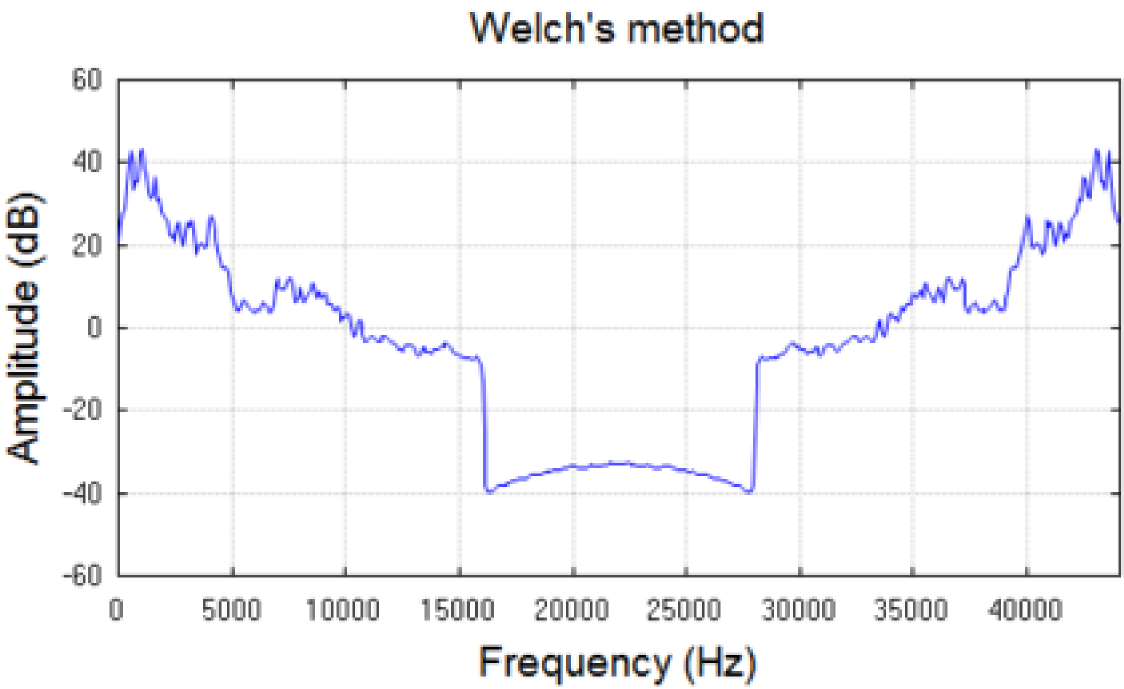
Through Fourier transform, as the name suggests is a technique of transforming or converting a signal such like seismic signal that we have used in the result of several frequency. Fast Fourier Transformation is a method of calculating the Fourier transform at a much faster rate (from N^2 form to NlogN time form).



6.3 Spectral Welch Density

Power spectral density primarily tells that what fraction or proportion of variance in the original frequency was produced by the given set of frequency that was breakdown by the Fast Fourier Transform. The Spectral welch Density is Power Spectral Density. Welch's method of computing said distribution.





CHAPTER 7

MACHINE LEARNING TECHNIQUE FOR PREDICTION OF EARTHQUAKE

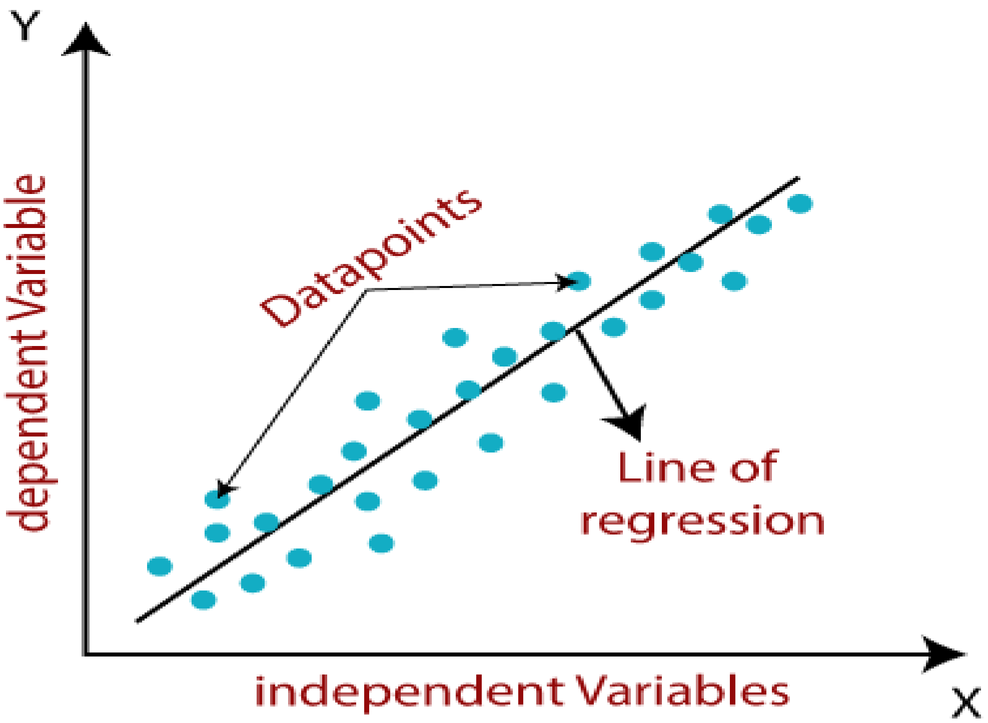
Various type of machine learning techniques are applied to acoustic data collected from laboratory micro earthquake simulation. In prediction process, six machine learning techniques including Linear Regression, Support Vector Machine, Random Forest Regression, Case Based Reasoning, XGBoost and Light Gradient Boosting Mechanism are separately applied and accuracies in the training and testing datasets were compared to pick out the best model. After the training of those techniques, the models are tested on more than 500 quantum of test data, and the performance is evaluated.

7.1 Linear Regression

It is a supervised learning based machine learning algorithm. It carries out a regression task. Centered on independent variables, regression models a desired prediction value. The Value is predicted in such regressor models, by establishing a relationship between the available observation of dependent and independent variables. In Linear regressor, the aim of the model is to find a linear relation. To execute this the model tries to draw, what is called a 'best fit line'. A best fit line is a line which aims to pass "as closely as possible from all the points observed in the data set. For, this it uses a mathematical function

$$m = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$b = \frac{\sum y - m\sum x}{n}$$



The above function essentially minimizes the sum of perpendicular distances between the line and all the points observed in the data.

Linear regression is used to estimate the value of a dependent variable (y) depending on a given independent variable (x). As a result, this regression method determines a linear relation between y (output) and x (input).

Y= θ1 + θ2*X

We are given the following instructions to follow while training the model:

θ1: intercept of y

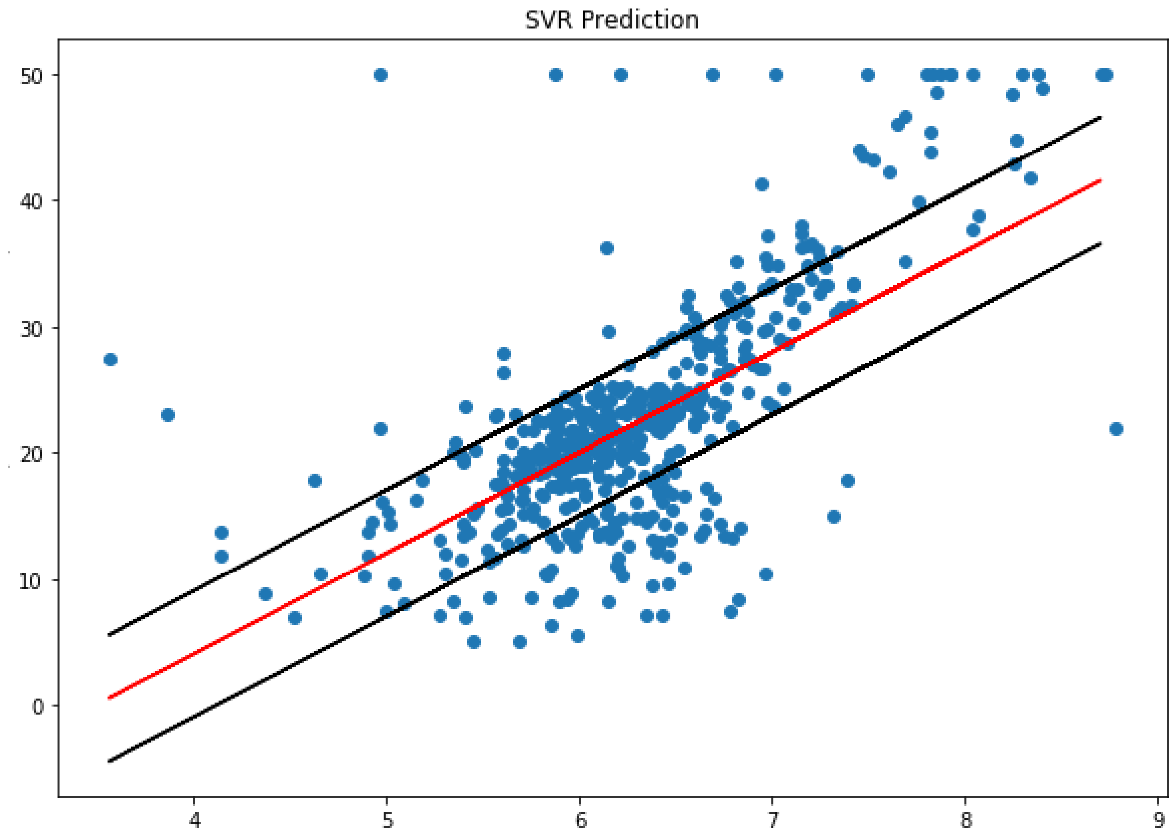
θ2: coefficient of x

It matches the best fit line after we find the best θ1 and θ2 values. So, when we actually use our model to simulate, it will predict the value of y based on the input value of x.

7.2 Support Vector Machine

It is a commonly supervised using algorithm that is used for both classification and regression problems. However, we had used it in Machine Learning for regression problems.

The Support Vector Machine algorithm's main aim is to find out a line that is best also called decision boundary for categorizing n number of dimensional space. That we can conveniently position data points in the best category in the future. This deciding boundary is called as hyperplane.

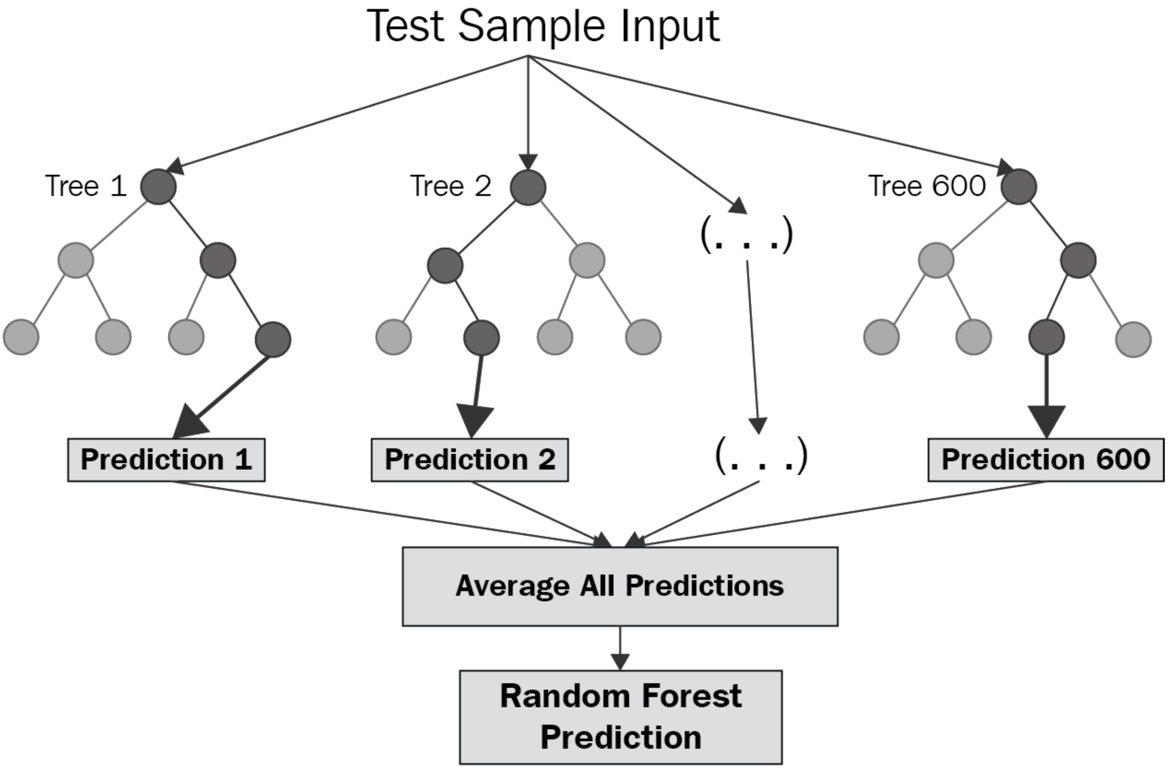


7.3 Random Forest Regression

It solves regression as well as classification problems by using ensemble methods (bagging). Any training phase, the model constructs n no. (where n is usually depends upon sample space, usually n is the square root of sample space).

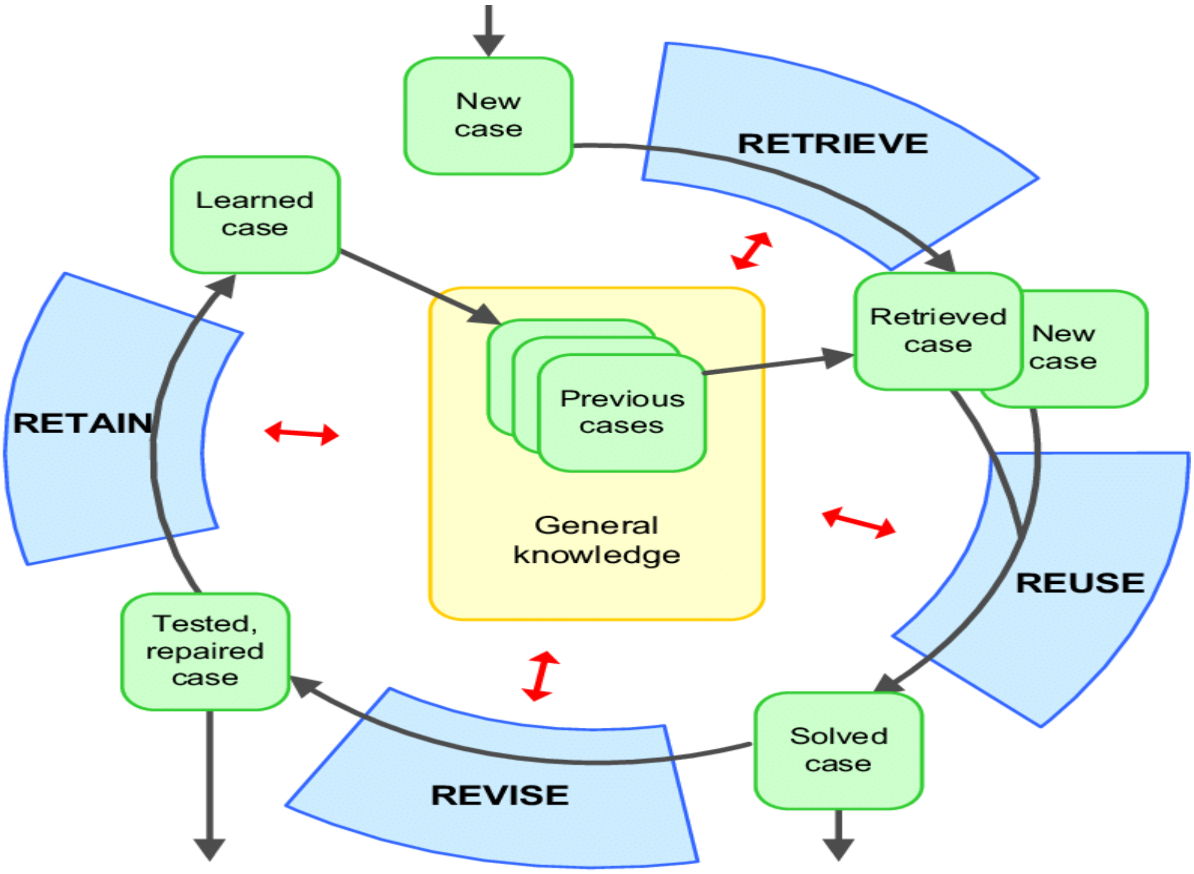
Random forest applied on the principle of 'wisdom of crowds' which states that a large number of differentiated models that is working like committee could perform outstandingly every set of the individual constituent models.

The explanation for this is that the trees guard each other from their own mistakes. A random forest functions as an estimator algorithm, aggregating the results of multiple decision trees and then producing the best possible outcome. In this case, 60 trees are selected for developing ensembles based on the concept of experimentation.



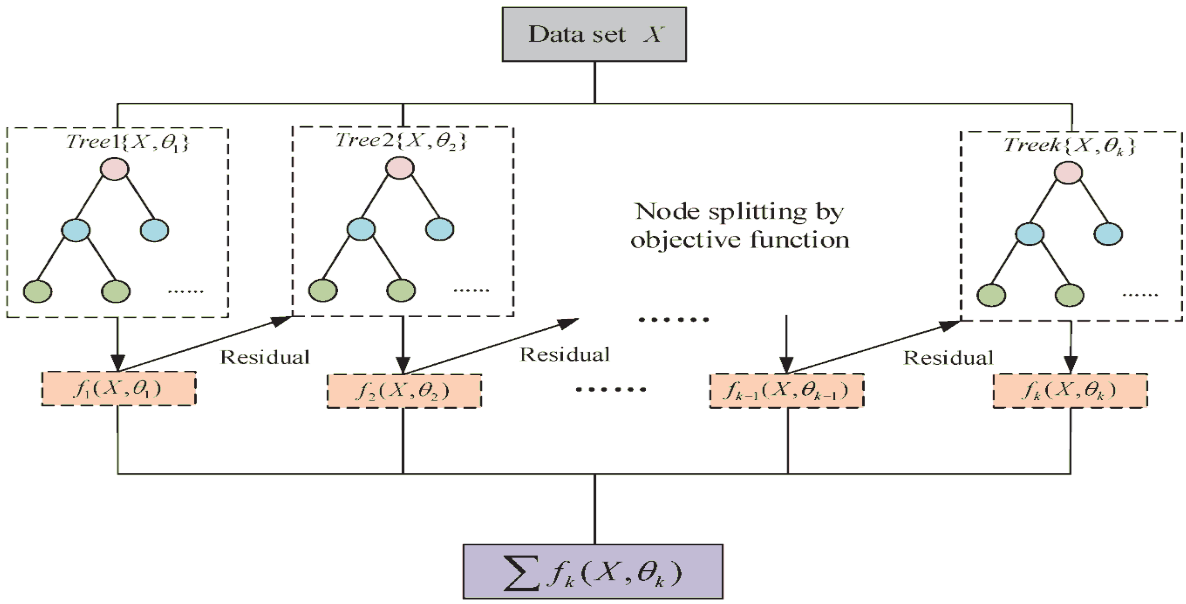
7.4 Case Based Reasoning

Case Based Reasoning (CBR) analyze a database of problem solutions to solve new problems. It saves problem-solving tuples or cases as complex symbolic definitions. When a new case emerges to classify, a Case based Reasoner can first search to see whether an equivalent training case exists. If one is detected, the case's corresponding solution is returned. If no equivalent case is detected, the Case Based Reasoner will look for training cases with similar elements to the current case. Conceptually, these testing cases may be considered of as the latest case's neighbours. If the cases are represented as graphs, this entails looking for subgraphs that are close to subgraphs in the new case. To suggest a solution for the current situation, the Case Based Reasoner attempts to merge the solutions of neighbouring training cases. If there are incompatibilities with the particular solutions, it could be important to go out and look for other solutions. To suggest a viable solution, the Case Based Reasoner can use background experience and problem-solving techniques.



7.5 XGBoost

XGBoost, also known as extreme gradient boosting, is a famous gradient boosting application (ensemble) that improves accuracy and makes it fast in sequential decision trees based machine learning algorithms. In boosting, trees are constructed in a sequence, with each successive tree attempting to reduce the errors of the previous tree. Each tree learns from the trees that came before it and updates the residual errors. As a result, the next tree in the series will benefit from a modified version of the residuals. It uses parallel tree boosting to solve a range of data science problems quickly and accurately.



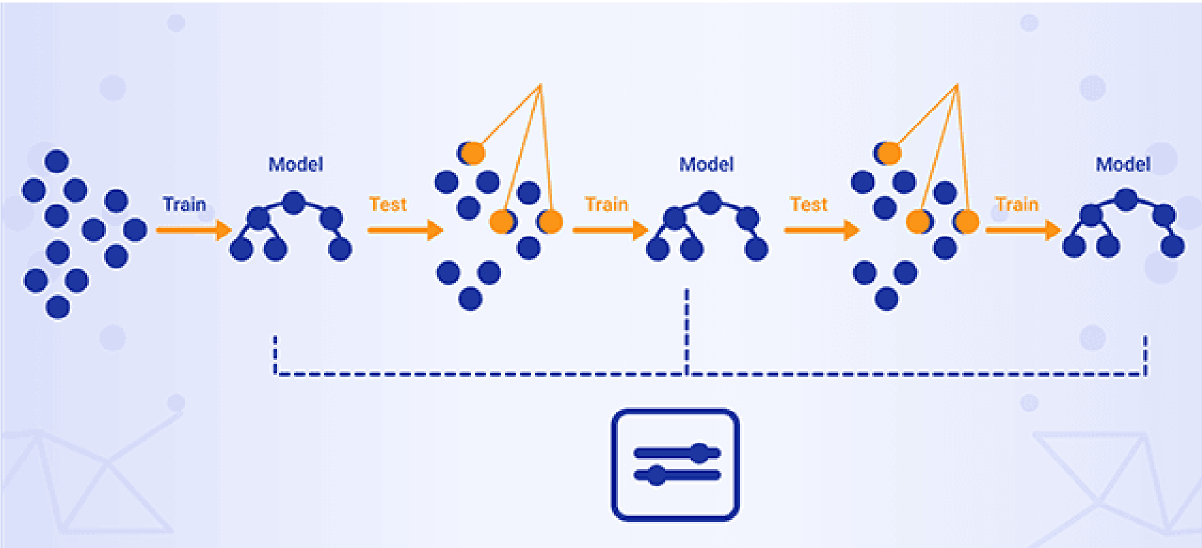
7.6 Light Gradient Boosting Mechanism (LGBM)

Light Gradient Boosting Mechanism corresponds to ensemble Machine Learning algorithm, used to solve regression predictive modelling problem.

Decision tree models are using to construct the ensembles. In ensemble construction, trees are added one by one to choose correct estimation errors made by previous models. It is a boosting algorithm, which is a kind of ensemble machine learning model.

Random arbitrary differentiable loss function and the gradient descent optimization algorithm are used to adapt models. Gradient boosting is named so, since loss of gradient is reduced, since the model is fitted like a neural network.

First folds for cross-validation are defined. Model Parameters were then defined. Run the model. During training for every fold, we validate using the validation set and we also predict using the present model for the test set. The ultimate results are going to be the typical over the all folds for the predictions done at each fold training.



CHAPTER 8

MODEL

There are many performance evaluator available that can be used for binary classification problems. The accuracy for prediction of earthquake is measured using the following measures:

8.1 Mean Absolute Error

Mean Absolute error, as the name suggests, is the mean of all the errors obtained in each of the regressor model's predictions and the actual observation. It can mathematically be expressed as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

8.2 Cross Validation Score (CV Score)

Cross-validation is a analysing method that is used to test models of machine learning on a limited set of sample data.

In this method, we had used single parameter known as k that indicates the number of sets of data into which we should divide the given sample of data. As a consequence, this methodology is usually known as k-fold cross-validation. Whenever there is new value for k used, it could be used in place of k in model's comparison, for e.g., k=15 resulting in 15-fold cross-validation.

Cross-validation mainly used for applied machine learning to estimate a machine learning model's capability on unseen data. That is to take a bit of segment to make assumption about the data in general,

The following is the general procedure:

1. Randomly shuffle the dataset.
2. Divide the data into k classes.
3. For every distinct group:
   a) Consider the group to be a holdout or evaluation data collection.
   b) Consider the remaining groups to be a testing data collection.
   c) Fix a model to the training data and then test it on the test data.
   d) Keep the test score and neglect the model.
4. By using the model's sample assessment scores, summarize the model's abilities.

8.3 Model Evaluation

We compared Cross Validation (CV) Scores of different machine learning models namely including Random Forest Regression, Linear Regression, Case Based Reasoning, Support Vector Machine, XGBoost and Light Gradient Boosting Mechanism by plotting a box plot against Mean absolute Errors of time to failure, as shown in fig:

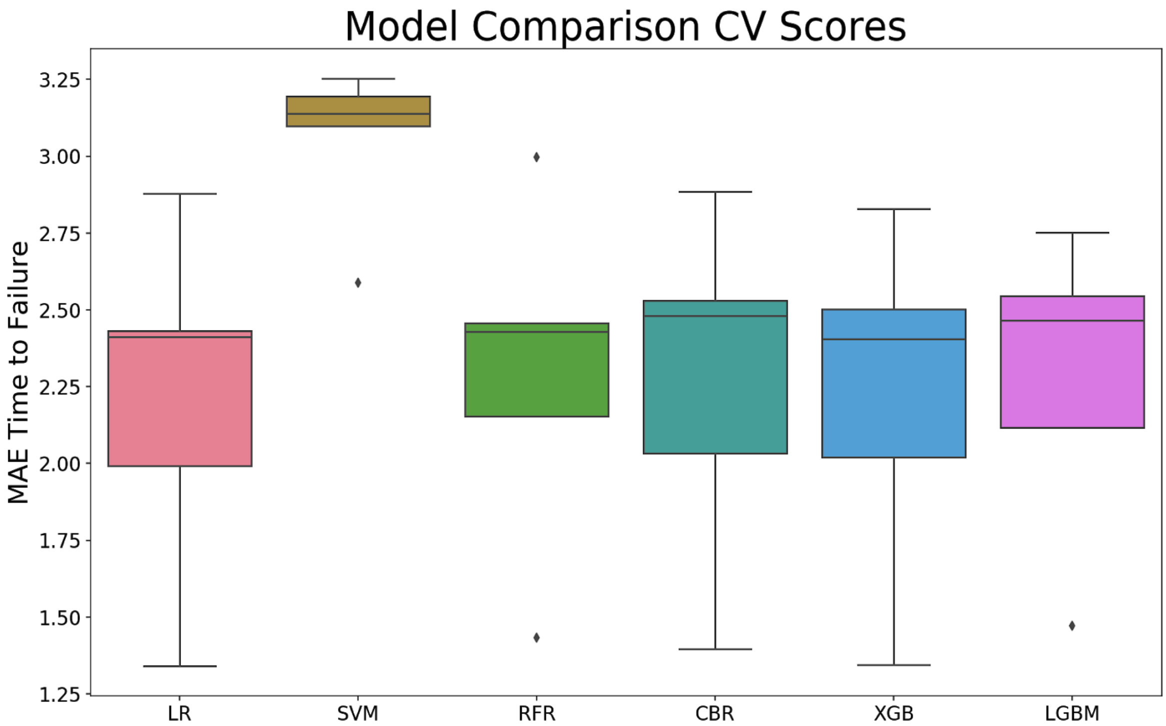
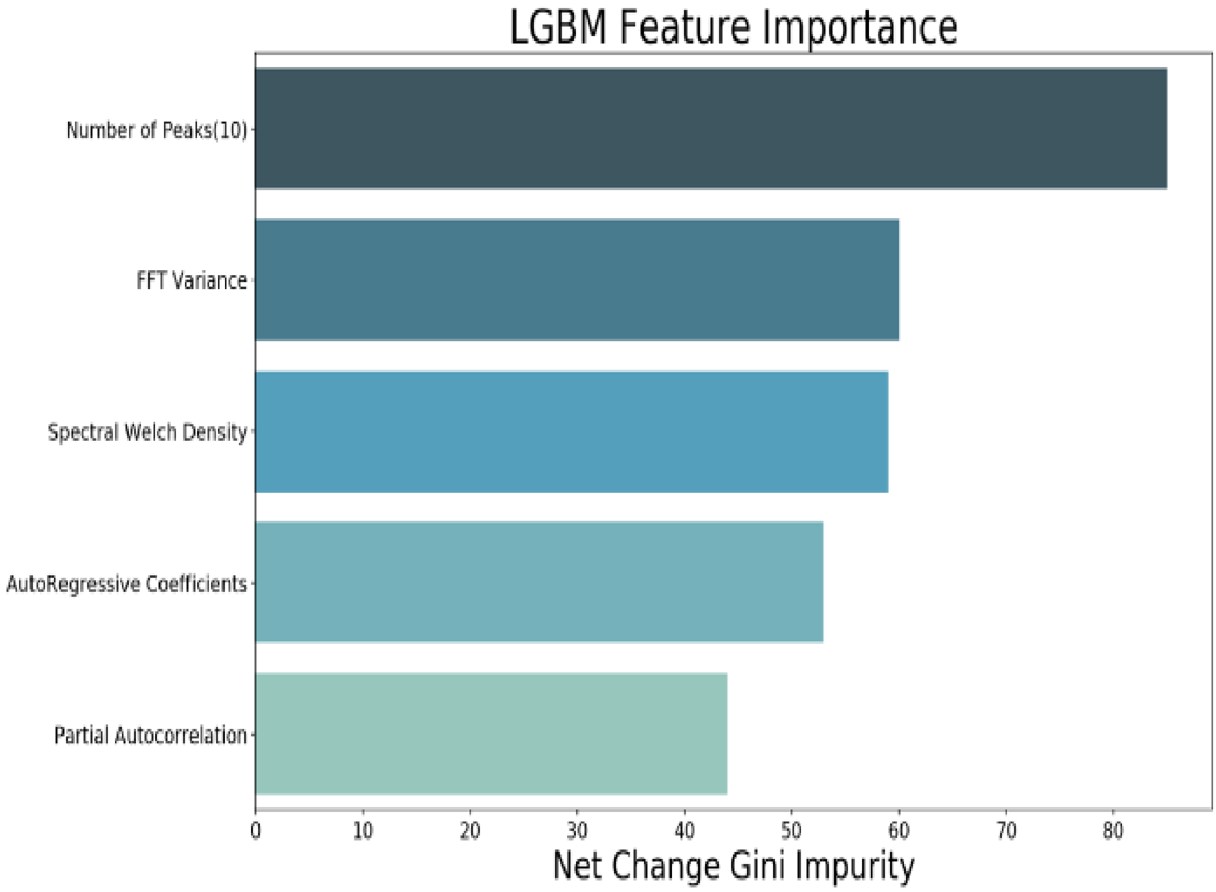


Fig: Boxplot of various ML models against MAE Time to failure

CHAPTER 9

RESULTS & DISCUSSION

After evaluating all the models and their CV scores, we concluded that the Light Gradient Boosting Model (LGBM) performs well as compared to its rest competitors, it has a fair balance between Mean Absolute Error (MAE) time to failure, and range of observations, and also has the least outliers. It is mean CV score for Mean Absolute Error time failure is approximately 2.4. The feature importance of the Light Gradient Boosting Model is also shown below:



ACKNOWLEDGEMENT

We are grateful to the almighty for establishing us to complete this B.Tech project. We are grateful to Dr. V. K. Minocha, HOD (Department of Civil Engineering), Delhi Technological University (Formerly Delhi College of Engineering), New Delhi and all other faculty members of our department, for their astute guidance, constant encouragement and sincere support for this project work. We owe a debt of gratitude to our guide, Dr. S. Anbu Kumar, Associate Professor, Department of Computer Engineering for incorporating in us the idea of a creative project, helping us in undertaking this project and also for being there whenever we needed her assistance. I also place on record, my sense of gratitude to one and all, who directly or indirectly have lent their helping hand in this venture. We feel proud and privileged in expressing my deep sense of gratitude to all those who have helped me in presenting this project. Last but never the least, we thank our parents and friends for always being with us, in every sense.

REFERENCES

[1] Los Alamos National Laboratory, Geophysics Group: Builds on initial work from Paul Johnson. B. Rouet-Leduc Bertrand Rouet-Leduc, and Claudia Hulbert prepared the data for the research.

[2] PennState, Department of Geosciences: Data are from experiments performed by Paul Johnson, Prof. Chris Marone, Jacques Riviere, and Chas Bolton.

[3] Department of Energy, Geosciences and Biosciences Division, Office of Science, Chemical Sciences, Basic Energy Sciences: The Geosciences core research.

[4] Purdue University, Department of Physics and Astronomy: This stemmed from the DOE Council workshop "Information is in the Noise: Signatures of Evolving Fracture and Fracture Networks" held March 2018 that was organized by Prof. Laura J. Pyrak-Nolte.