HALMSTAD UNIVERSITY

DATA MINING PROJECT REPORT

Project on Vehicle Dataset

Abdul Basit
Chandraprakash

Table of Contents

Abstract

In this Data Mining project, we got an opportunity to work on real world dataset and problem formulation. In which the dataset was provided which is based on vehicle information such as electric, hybrid &diesel mode vehicle use in different countries and there are total 74 features related to vehicle was present in the dataset. We have to analyze the dataset and generate challenging problems from it and then solve those challenging problem using machine learning techniques (Supervised and Unsupervised).

Introduction

Data mining is a process to apply a set of methods for data analysis, exploration and discovering the problem solution in a large and complex dataset. In this project, we have provided dataset related to vehicle information that operates in electric, diesel and hybrid modes and countries or geographical area in which the vehicles are used. The names of the countries are encrypted with numbers. This dataset contains total 74 features and 861916 samples corresponding to 4031 vehicles, however, the measurement of each vehicle are not updated regularly. The challenging task is to analyze the dataset and finding the problems and then solved by performing data mining process and machine learning techniques. After analyzing the dataset, we have found the below two problems and applied the data mining methods in which data preprocessing is one of the most important and challenging tasks are performed which includes data cleaning, data exploration and etc, then applied machine learning models which are used for finding the best optimal results for those problems.

Problem Statement

Supervised

a) Predicting the Operation time using Regression Models.

b) Prediction of multiclass classification feature ‘Battery Supplier’.

Unsupervised

Anomaly detection of State of Health of Battery and evaluation with different model on prediction of battery health.

Methodology

The following steps are performed to achieve the goals which are listed below ;

Data Collection :

Dataset was provided from the University and which was loaded and changed it into an appropriate format. Imported all the libraries which were needed to perform sampling, visualization, data modeling, data evaluation steps etc. Discovered the data to get insight on it, checked the data counts, statistical summary of attributes, break down the data based on target goal, and taken the features which are correlated to target variable for the prediction values.

Data Pre-Processing :

This is one of the most important processes, i.e. cleaning the raw data by removing duplicate samples, treating missing values, outlier removing and prepared for modeling and evaluation of data. When the data is gathered, it is collected in a raw format and this data is imbalanced and not feasible for directly analysis purpose, so certain steps are executed to prepare the raw data into clean data which can be analyzed and used for modeling purpose.

Machine Learning algorithms could not give good result with missing features, so it was handled by inserting the median values and removing the null data.

After loading the raw dataset, found that there are multiple duplicate and wrong values present in the dataset like a particular country should be lie in specific geographical area but instead of this it lie in different geographical areas. For an instance, the ‘Country’ 3 lie in geographical areas ‘1,2,3,5,6’. Other features also have wrong values like city 4 lies in different countries ’27,26,22,21,1,6’ and also raw dataset contains lot of null values.

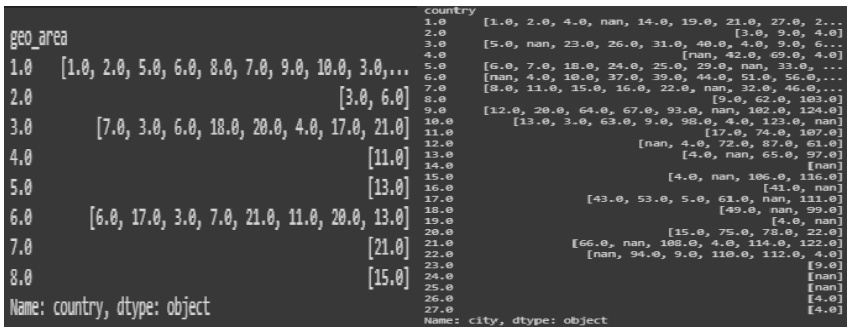


Figure 1 : wrong values in geo_area and in country .

a) While doing the data cleansing, we have also found that feature “Sample_Id ” is not unique, however, it should be unique for all recorded vehicle id. Lots of duplicate entries were there which are approximately 84449.

```
[ ]  1 duplocate_sample_id=raw_data.duplicated("sample_id").sum()
     2 duplocate_sample_id

     84449
```

Figure 2: Duplicate values in sample_id

b) For duplicate Sample_id, some features have different values, for instance, “Battery_Version”, “Battery_replacement_date”, “Battery_Genreation” and “Battery_Supplier” have different values, however, the sample id is same.

Sample_I	SEND_TIME	Vehicle_ID	Operation_D	State_Of_Healt	Total_Time	Electric_N	Hybrid_	Total_Dista	Battery_Ve	Battery_Replacen	Battery_Ge	Country	Mounted_E	Engine_T	IS_Battery	Battery_Supplie
7110248	2011-10-26 16:13	2928	507.65						0	2011-09-01	nan	1.0	3.0	2.0	0	nan
7110772	2011-10-26 16:28	2005	7692.6100000	86.37	739.1100000(	271.36	7661.81	7933.17	0	2011-09-01	nan	1.0	3.0	2.0	0	nan
7110787	2011-10-26 16:29	2725	2403.9700000	28.99	251.07	70.0	2396.14	2466.14	0	2011-09-01	nan	1.0	3.0	2.0	0	nan
7111023	2011-10-26 16:35	2005	2425.4	28.93	255.84	67.5	2417.55	2485.05	0	2011-09-01	nan	1.0	3.0	2.0	0	nan
7111222	2011-10-26 16:41	534	0.0						0	2009-04-01	nan	1.0	3.0	2.0	0	nan
7111222	2011-10-26 16:41	534	0.0						1	2016-05-31	3.0	1.0	3.0	2.0	0	3.0
7111230	2011-10-26 16:41	49	4850.35	54.65	461.4500000(	98.18	4838.12	4937.43	0	2011-09-01	nan	1.0	3.0	2.0	0	nan
7111230	2011-10-26 16:41	49	4850.35	54.65	461.4500000(	98.18	4838.12	4937.43	1	2016-08-24	3.0	1.0	3.0	2.0	0	3.0
7111671	2011-10-26 16:54	1724	4948.4000000	56.46	462.7900000(	219.7	4921.96	5141.99	1	2016-10-10	3.0	1.0	3.0	2.0	0	3.0
7111671	2011-10-26 16:54	1724	4948.4000000	56.46	462.7900000(	219.7	4921.96	5141.99	0	2011-09-01	nan	1.0	3.0	2.0	0	nan

Figure 3: Different values in some features for duplicate sample_id

c) The feature “Is_Pattern_available” gives the information of data logged or not logged. The numeric value ‘0’ shows the data is not logged whereas the value ‘1’ shows the data is logged, so taken the data which has value ‘1’ and got 40580 records has the value ‘1’.

```
[ ]  1 cleansed_data_on_pattern=raw_data[(raw_data["is_pattern_available"]==1) ]
     2 cleansed_data_on_pattern.shape

     (40580, 76)
```

Figure 4: Data cleasingwrt “Is_Pattern_available”

d) After selecting the logged value which is equal to 1, found that there are 106 duplicate sample id still exists, which are removed by considering the features ‘Battery_Version’and ‘Is_Battery_Changed’. We have taken these two features because the feature ‘Is_Battery_changed’ represent battery is changed or not if it turns to ‘1’ means the battery is changed and‘0’ mean not changed, therefore, if the ‘Is_battery_changed’ is equal to ‘1’ then ‘Battery_Version’ should not be equaled to ‘0’ as ‘Battery_Version’ represent number of times battery changed and vice versa.We have also found while narrow down the data that there are 23 records which

values are ‘0’ in feature ‘Battery_Version’ and ‘1’ in ‘Is_Battery_Changed’ which are wrong therefore it was removed.

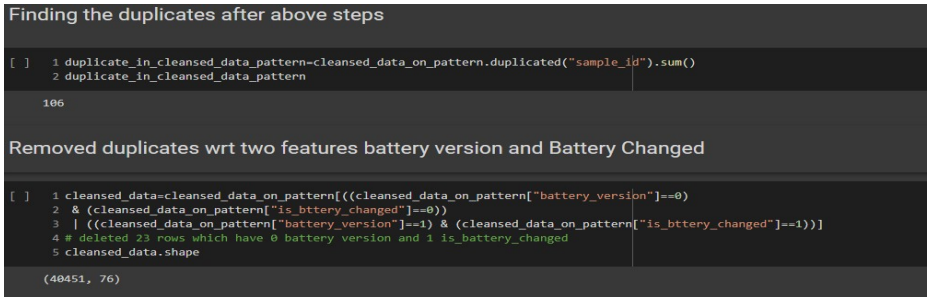


Figure 5: Duplicate data handling

e) The below graph shows the missing values present in all the features in which the maximum 93% values are missing for active_charge feature. We have used only those features which have lesser than 40% missing values and removed the rest. After removing, the shape of the data set is (40451,42).

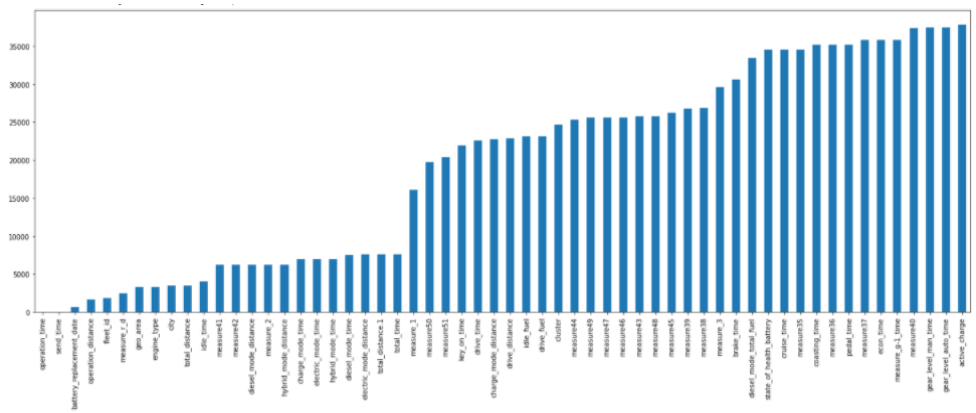


Figure 6: Missing values(%) in features

f) After cleaning the raw data, splitted the data into numeric and categorical features. After that, checked the outliers on numerical data, the below box plots show the outliers :

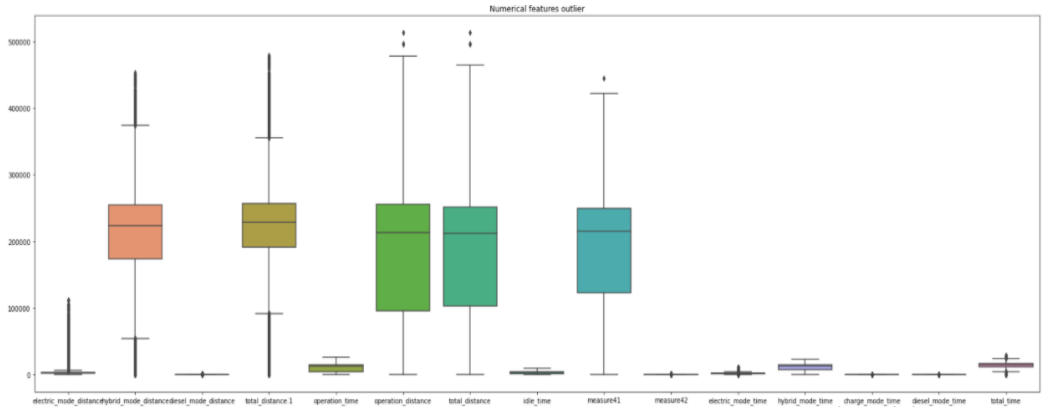


Figure 7: Numerical Feature Outlier

Here, we could see, the outlier is not showing in time features because the scale of the distance features is larger, so, we divide numeric dataset into further time or distance features for checking the outliers. Shown in below figures.

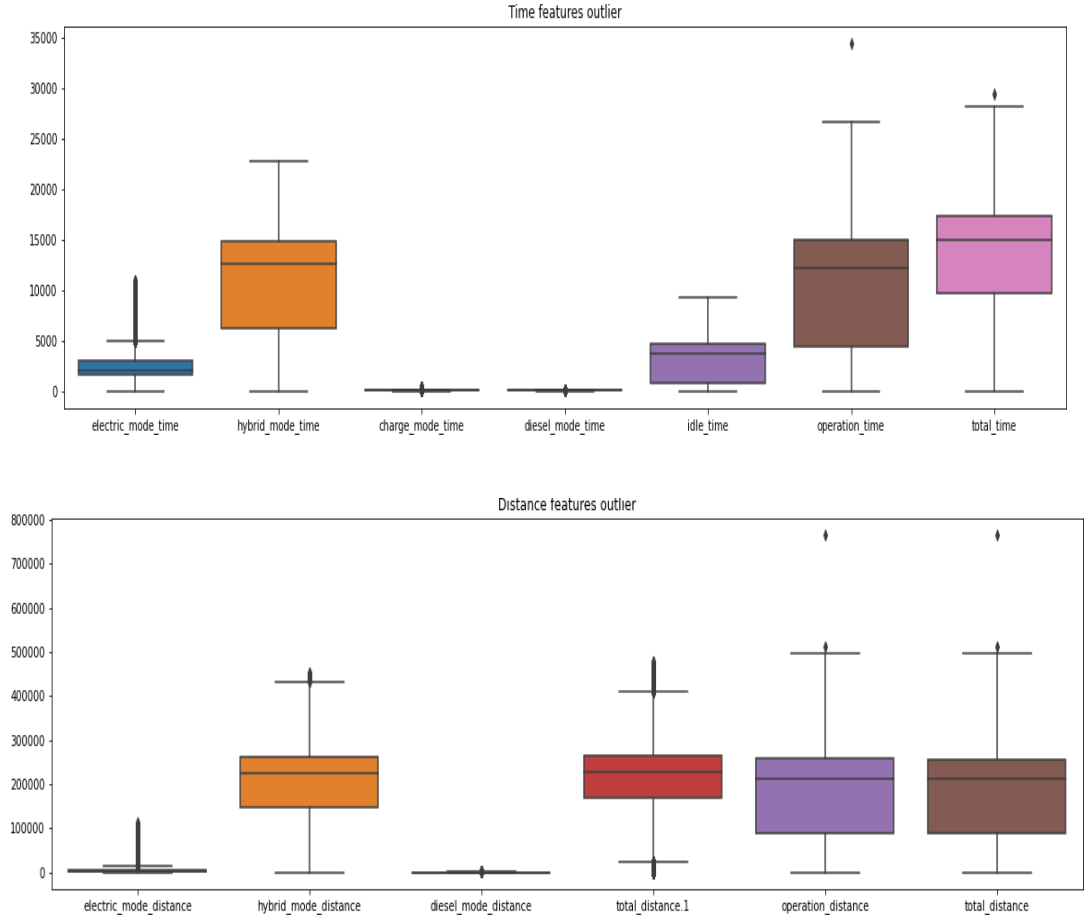


Figure 8: Time & Distance Feature Outlier

g) After removing the outliers from distance and time features, we checked the distribution of the numeric data, and verified whether it is normally distributed or skewed. The distribution says that which statistical method can be applied on dataset for the imputation. The below graph shows the numeric distribution

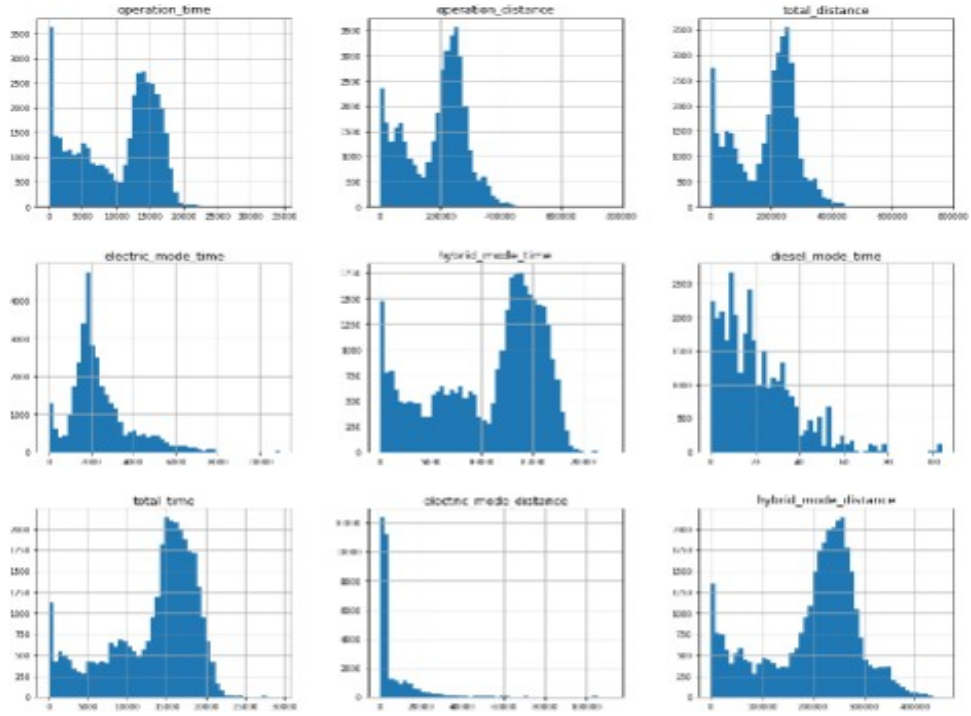


Figure 9: Features Distribution

From above figure, it shows that data is not normally distributed it skewed toward positive or negative. In this case, taking the mean for imputation is not a good idea because mean takes every data points. If we have a outliers than it effect the distribution of the dataset. It can drag the mean up and down. So, in this case, median would be the best choice for imputation.
Moreover, for categorical features used mode to fill null values.

For numerical feature selection, we perform Pearson Correlation by using heat map and analyze that features are highly correlate with each other, so, we apply PCA on features because PCA works well when we have a strongly correlated features. For categorical features, we also apply PCA. The below figure shows the numerical feature correlation heatmap.

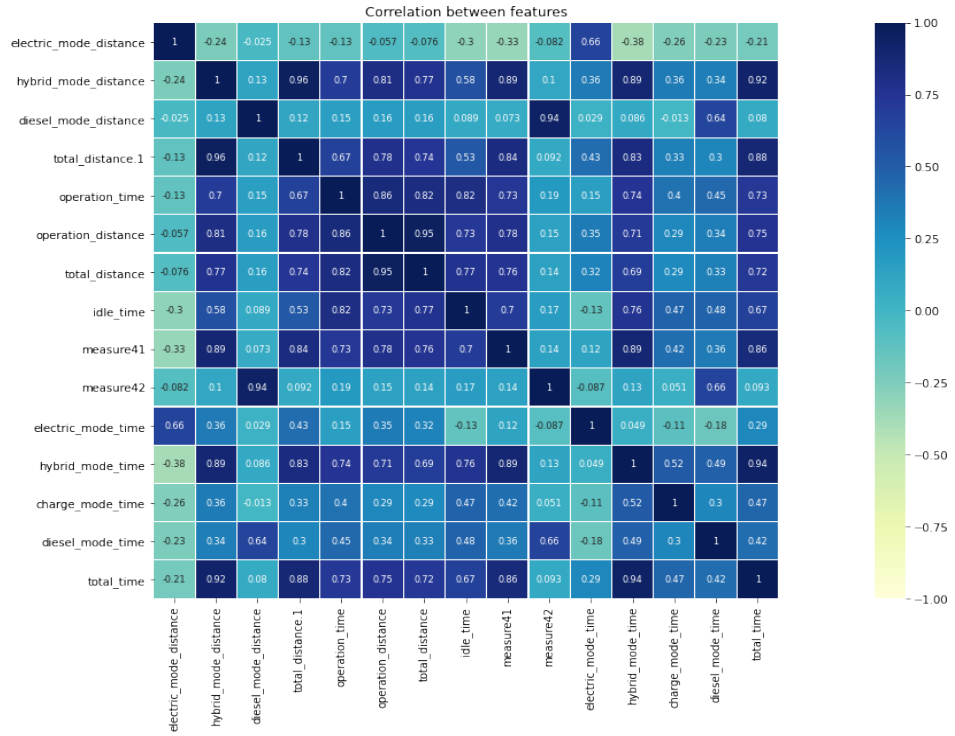


Figure 10: Correlation between features

h) Feature Scaling: Normalizing the numeric features, which is then used when the features are in different ranges. If we have a feature with a large range, then it will affect the model. So, it is important to take all features in a same range to perform the model better.

When we have a multiple features and each feature have different range, magnitude and units then we need to perform feature scaling. For an example there are some features in the data set is distance features in kilometers and time features in hour, therefore, we need feature scaling technique to train our machine learning model.

In Normalization, values are shifted and rescaled as resulting the data has range between 0 and 1 and we also called this technique Min-Max scaling. Normalization technique is good when we have a non-gaussian distributed data. Formula for normalization.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardization is another technique for scaling, In this technique, the values are centered around the mean with a unit standard deviation. Standardization works well when we have a gaussian distributed data. Formula for standardization.

$$X' = \frac{X - \mu}{\sigma}$$

The below figure shows the normalization on numeric features.

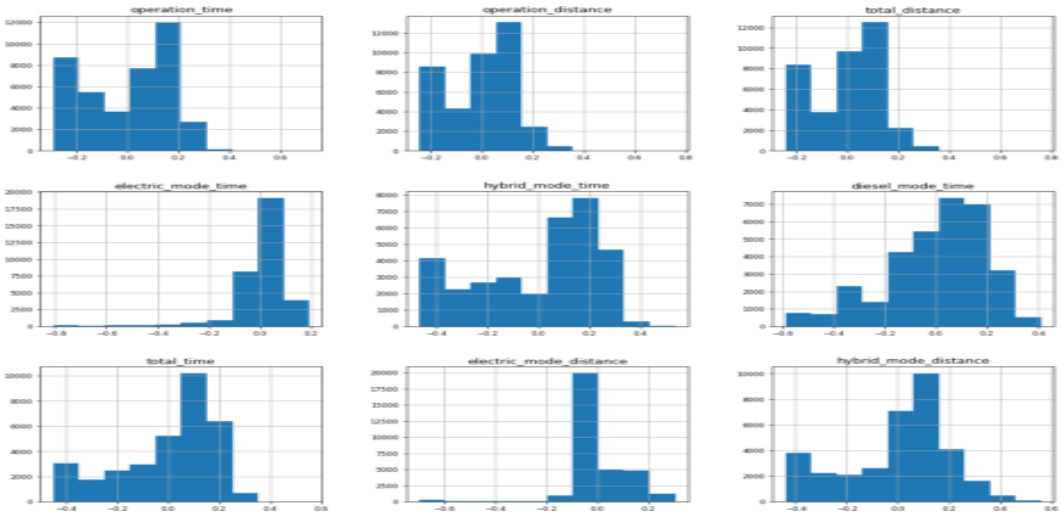


Figure 11: Normalization numeric features

i) After normalization of the data, we applied PCA (principle components analysis) on both numerical and categorical features after selecting the target feature.

j) Used SMOTE(synthetic minority oversampling technique) for handling the class imbalanced data in classification task. We are doing oversampling on minority class to make the prediction better.

Exploratory Analysis

a) Distance by Vehicle mode with respect to countries:

Mostly, Hybrid mode vehicles are used based on distance covered. For an example, country 2 covered distance 250000km in hybrid mode and in electric mode country 4 covered distance of 35000km. none of the countries switched to diesel mode for longer distances.

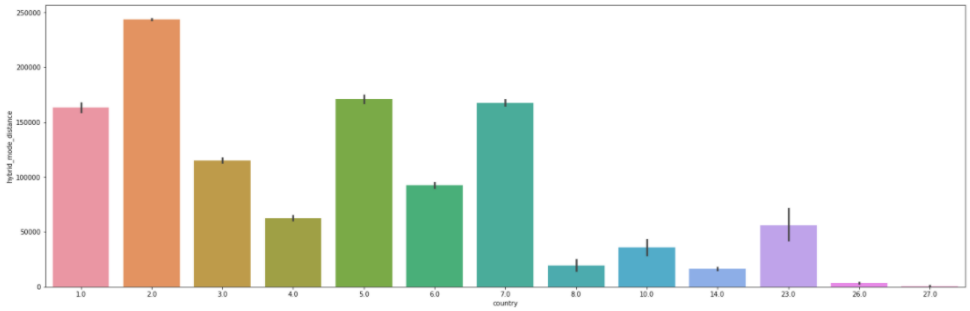


Fig 12: Hybrid_mode_distance

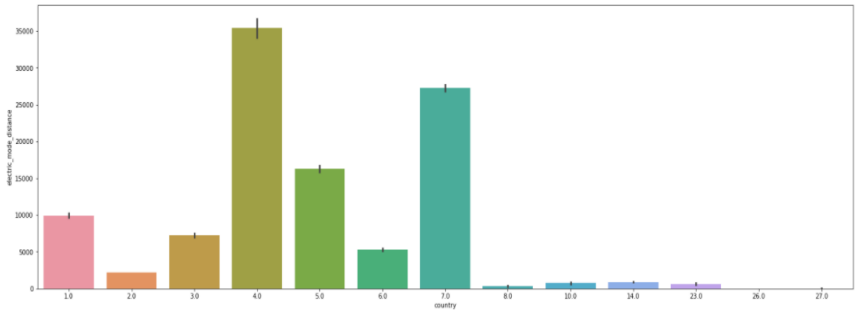


Fig 13: Electric mode distance

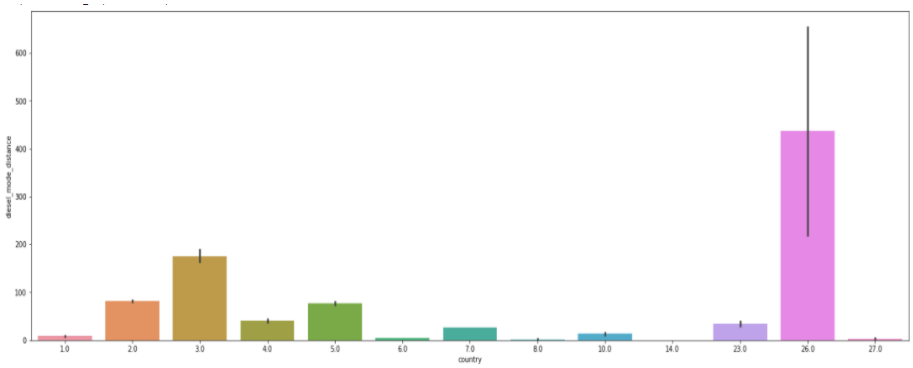


Fig 14: Diesel_mode_distance

b) Visualization of Engine type and Battery supplier relationship.
There are two engine types have in this dataset. Engine type 1 used batteries supplied by the supplier 1,2 and 4 whereas engine type 2 used batteries supplied by supplier 2.Before cleaning the data, we could see the engine type 3 & engine type 4 are exits, however, after cleaning, we could see, battery supplier 1,2& 4 data are available for engine 1 & 2.

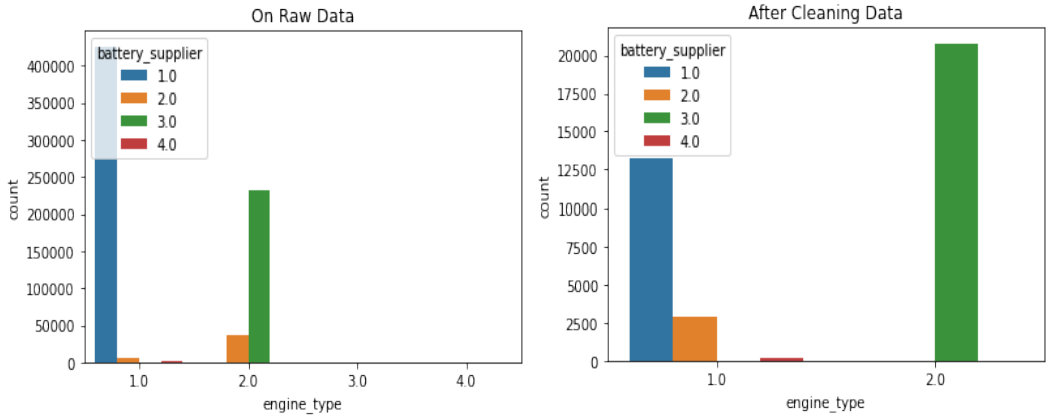


Fig 15: Engine Type

c) Similarly, In the battery supplier on raw data, battery generations are exits for all supplier, however, the counts are more for battery generation 1 and 3 in supplier 1 and 3 respectively. On other hand, after cleaning the data, the counts for battery generation 1 is decreased while other are nearly same counts before and after cleaning the data.

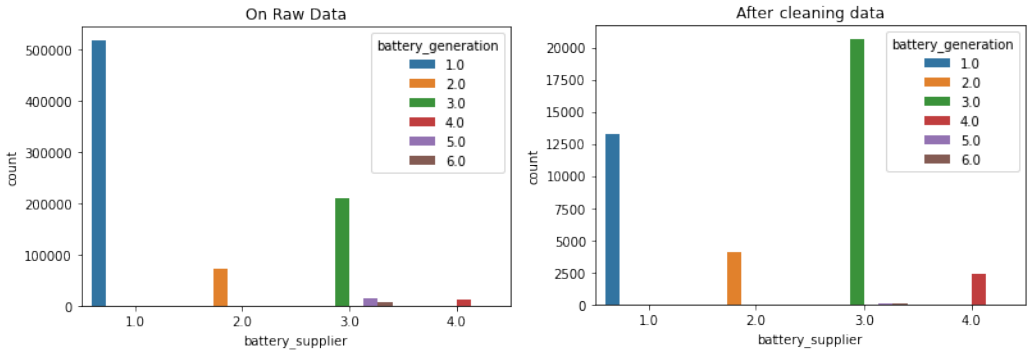


Fig 16: Buttery Supplier

d) The below graph shows the relation between Electric mode distance and country wise for 'Battery_Supplier','Battery_Generation', 'Battery_Version', 'Engine_Type', 'Emission_ Level'.

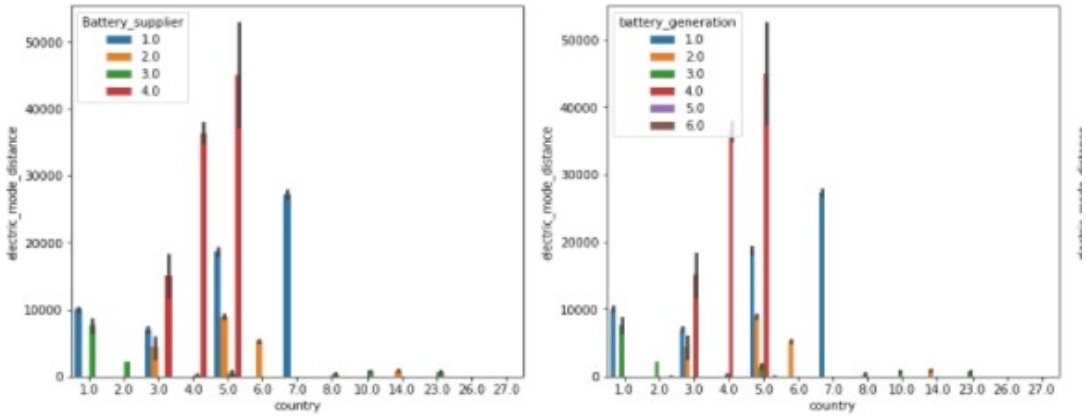


Fig 17: Country

In the first graph shows that the battery supplier 4 for country 5 is better than the rest of the other suppliers based on the distance covered by electric mode vehicles. In Electric mode distance, mostly battery generation 4 is used which is supplied by supplier 4 in countries 3, 4 & 5.

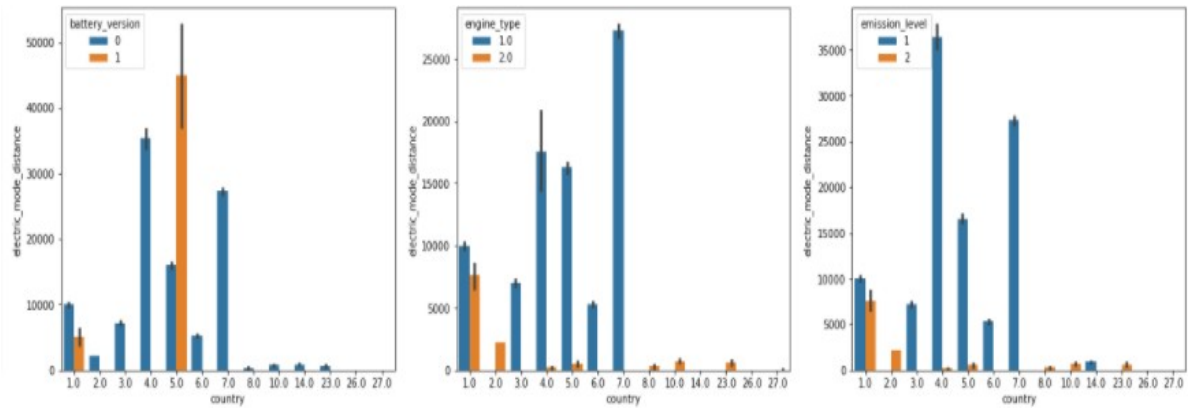


Fig 18: Relations between Country and Electric mode distance for Battery_Version, Engine_Type and Emission_Level

In the above graph 'battery version', which gives the information of number of times battery changed, so the maximum number of battery changed was in country 5 and the engine type 1 is used in most of the countries and maximum for country 7, moreover, the emission level'1'is majorly use in all the countries and highest for country 4.

Machine Learning

Supervised & Un -supervised Machine Learning

In Supervised Machine learning, we have used Classification algorithms i.e. Logistic Regression, Support Vector Classifier and in Regression algorithms, have used Linear Regression and Support Vector Regression. While, K- means clustering method is used in Un-supervised machine learning in the dataset for predicting the problem statements. The below methods are performed for data preprocessing, feature selection, data modeling and for finding the best classifier:

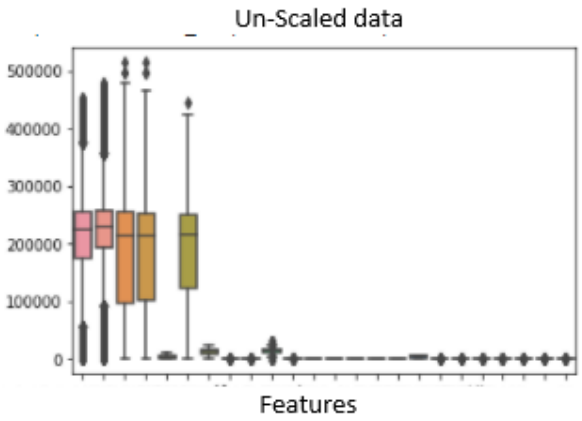
- Splitting the data into Training and Testing
- Used PCA
- SMOTE for oversampling
- Training and Testing the model
- Evaluation of Model

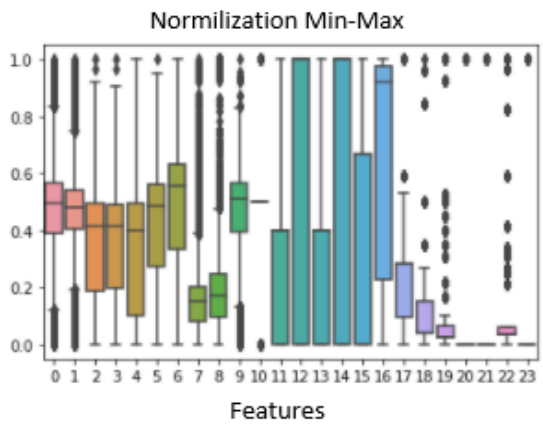
Supervised Task

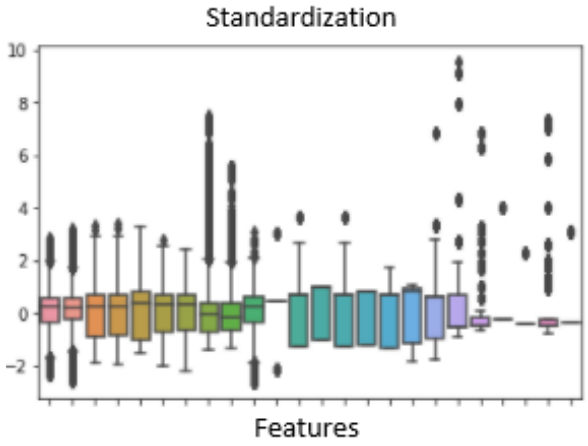
Problem Statement

a) Predicting the Operation time using Regression models

- Finding correlation on all features with respect to 'Operation time' and applied threshold of greater than 0.25 and less than -0.25.
- Below boxplot shows the comparison between unscaled or scaled data.







After applying the feature scaling which is explained in data-preprocessing section (h), we trained our models, the results shown below.

Result:

We have checked by performing the below algorithms. Results shows that when we have a normalized data, we have a less error as compare to standardized and original data. Because our data is not a gaussian distributed, so after training the models, the results shows the Min-Max normalization perform better and give less error in KNN model.

Model : Support Vector Regression			
S.No	Scaling	RMSE	R2
1	Original	3016.634642	0.743569
2	Normalized	2494.941269	0.824594
3	Standardized	2501.445159	0.823678

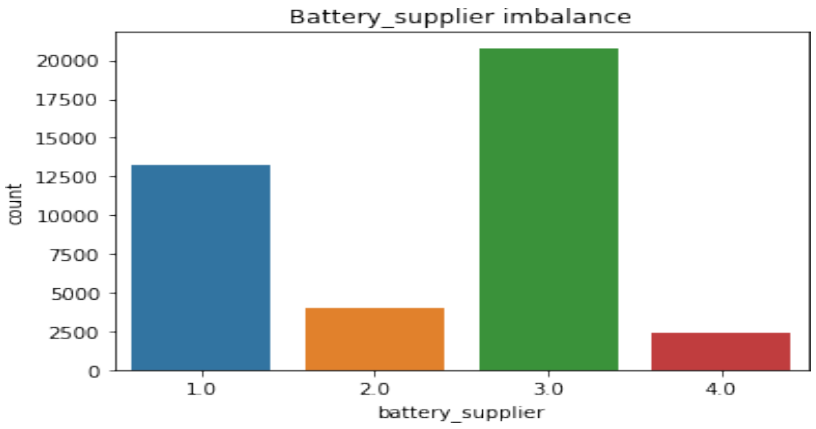
Model : KNN			
S.No	Scaling	RMSE	R2
1	Original	734.603734	0.763569
2	Normalized	504.605211	0.854594
3	Standardized	512.061733	0.853678

b)

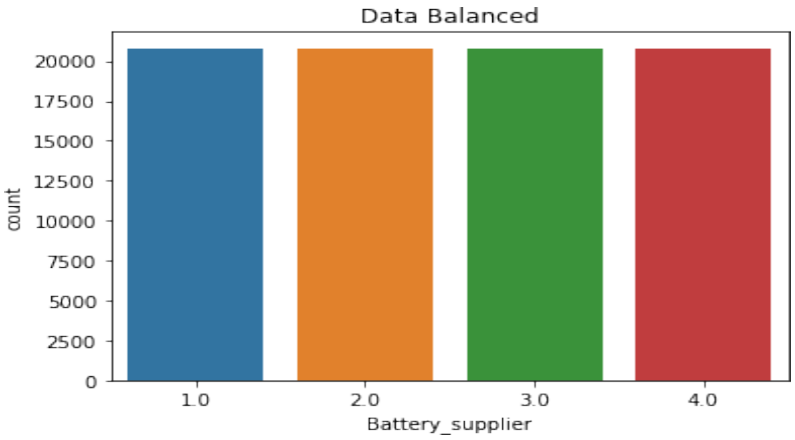
B) Prediction of multiclass classification feature 'Battery Supplier'.

- Finding correlation on all features with respect to 'Battery Supplier' and applied thresholdof greater than 0.25 and less than -0.25 .
- Applied SMOTE on Battery Supplier for balancing the data.

- Before SMOTE (Imbalanced)



- After SMOTE (Balanced)



- **Result without PCA and SMOTE**: After performing the below algorithms in this problem statement, found that Support Vector Machine has more accuracy than the Logistic Regression without PCA and Class Imbalance data.

S.no	**Model**	**Accuracy**
1	Logistic Regression	90.52 %
2	Support Vector Machine	95.31%

- **Result with PCA and SMOTE** : After performing the below algorithms, found that Support Vector Machine has more accuracy than Logistic Regression with PCA and SMOTE.

S.No.	**PCA**	**Logistic Regression Accuracy**	**Support Vector Machine Accuracy**
1	2	63.66%	89.12%
2	5	66.11%	89.86%
3	10	71.0%	92.22%

Unsupervised Task

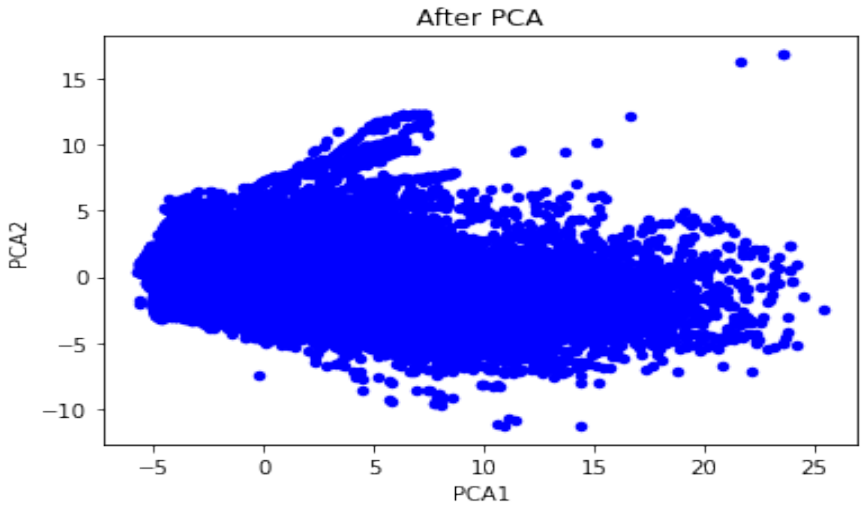
Problem statement

Anomaly detection of State of Health of Battery and evaluation with different model on prediction of battery health.

- In unsupervised algorithm, we have used the same cleaned dataset and applied different unsupervised clustering method i.e. HDBSCAN, Isolation forest for finding outliers on State of Health of Battery and then used for prediction of battery health and checked the error before and after anomaly.

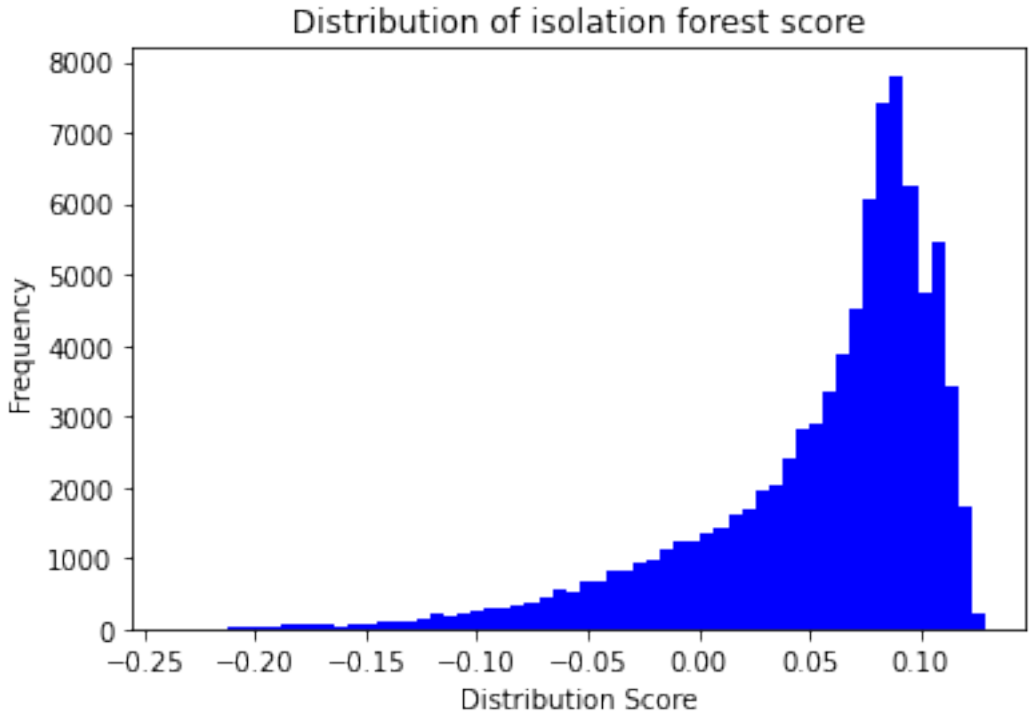
  It is very important to identify the anomalies and solved it because it is one of the serious problems in machine learning as it can cause over fitting of the data and predict bad results.

- We have used PCA for dimensional reduction to identify the anomalies in the data. We have reduced the dimension and then plotted the data from which, got the maximum variance, also from the plot got the clusters according to the arrangement of the data and there are some data which easily detect as outliers.

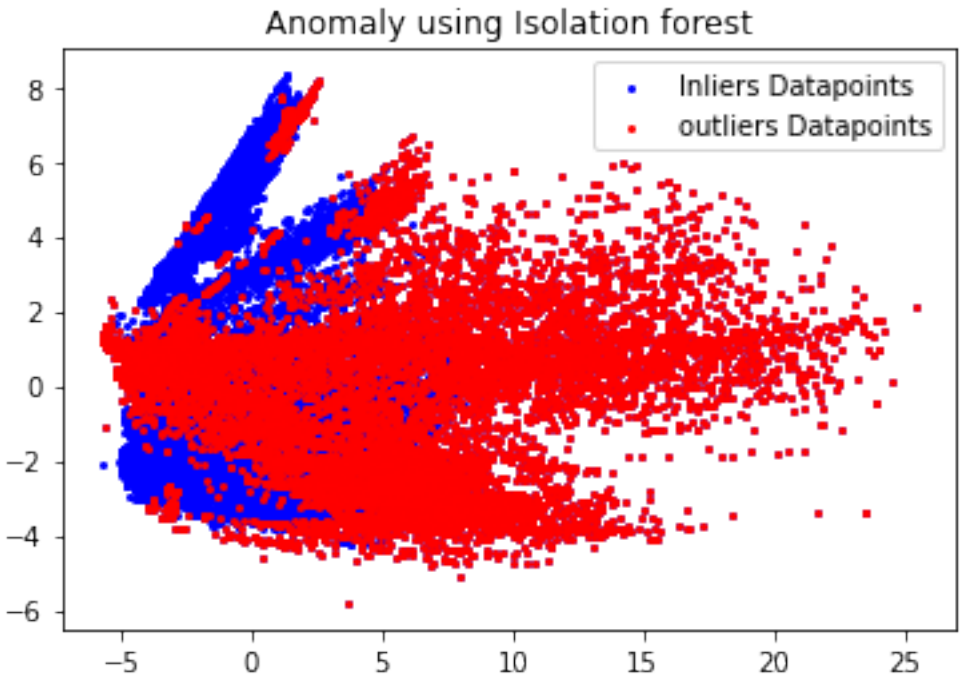


Isolation Forest for outlier detection

- Isolation forest is a unsupervised, tree based algorithm, It is built on the theory of decision tree and random forest, and it's a part of ensemble model class, available in sklearn.
- It is one of the fastest as well as consume less memory than the other anomaly algorithms.
- Isolation forest will train on the dataset and predict the data points into anomalies by marking -1 and 1. Also, we will find the scores, which indicates, lower values lesser than 0 observes anomalies and on other hand the values above 0 is considered as normal.
- we have used the same cleaned dataset and applied isolation forest algorithm.
  In the below fig, we could see that the negative scores indicates anomaly and the positive are normal.

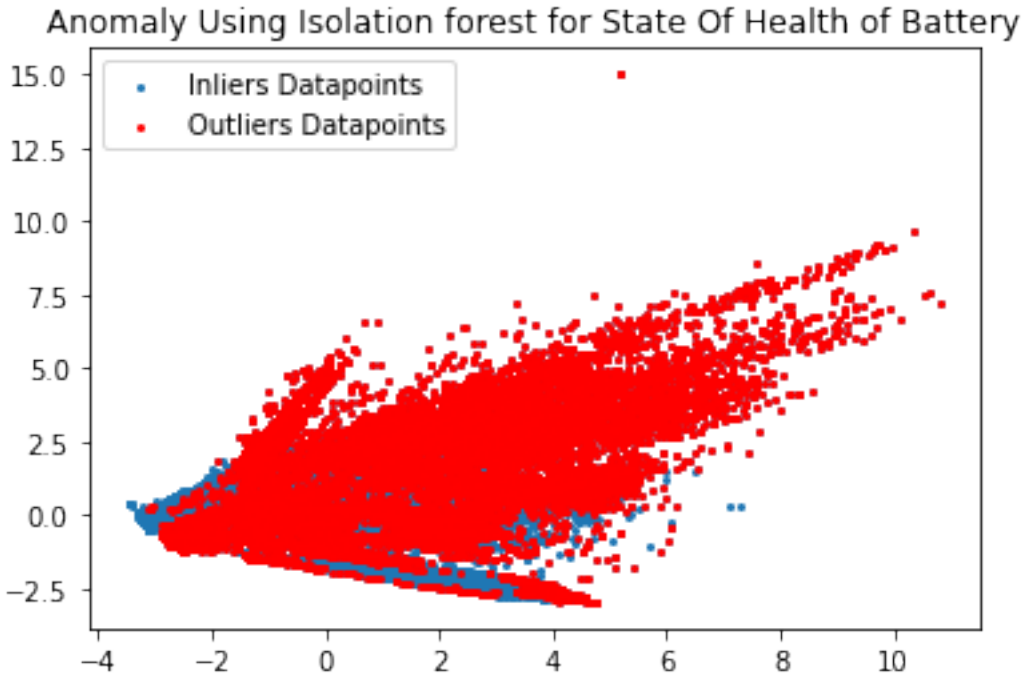


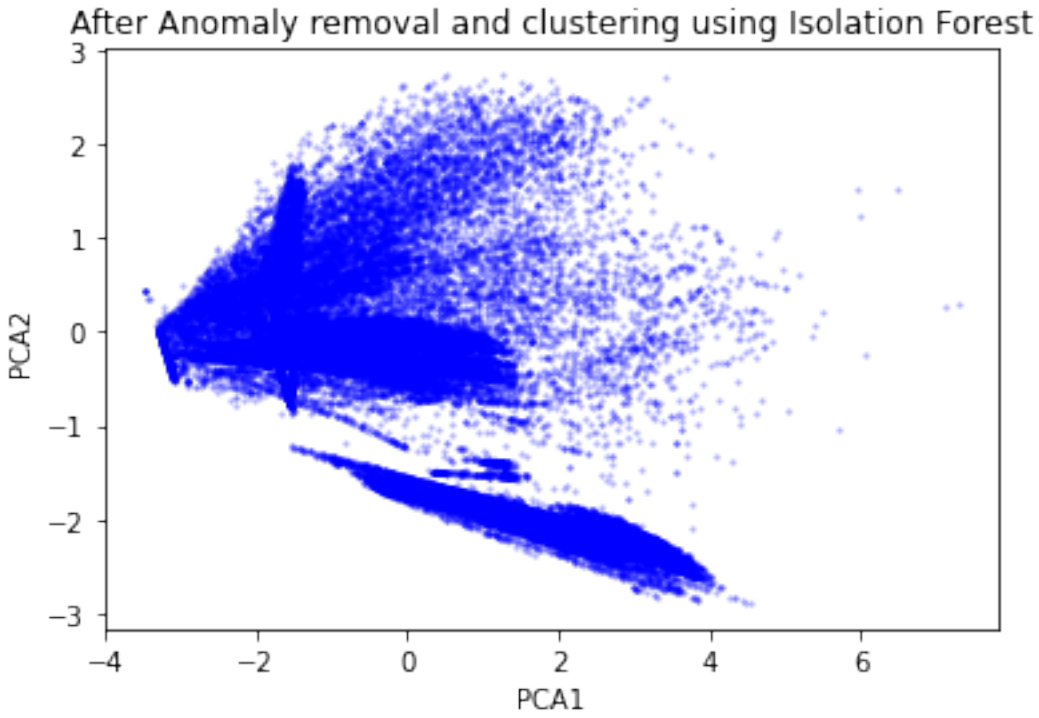
- The below figure shows the Anomaly(outliers) and normal(Inliers) data points using Isolation Forest.



Prediction of Battery Health using Isolation Forest:

For the prediction of state of health of battery using Isolation forest, we will train the data and predict the anomalies which lies outside the normal data points, It means we will take which are lesser than zero points. After that, all the outliers are removed and clustering using the Isolation Forest.



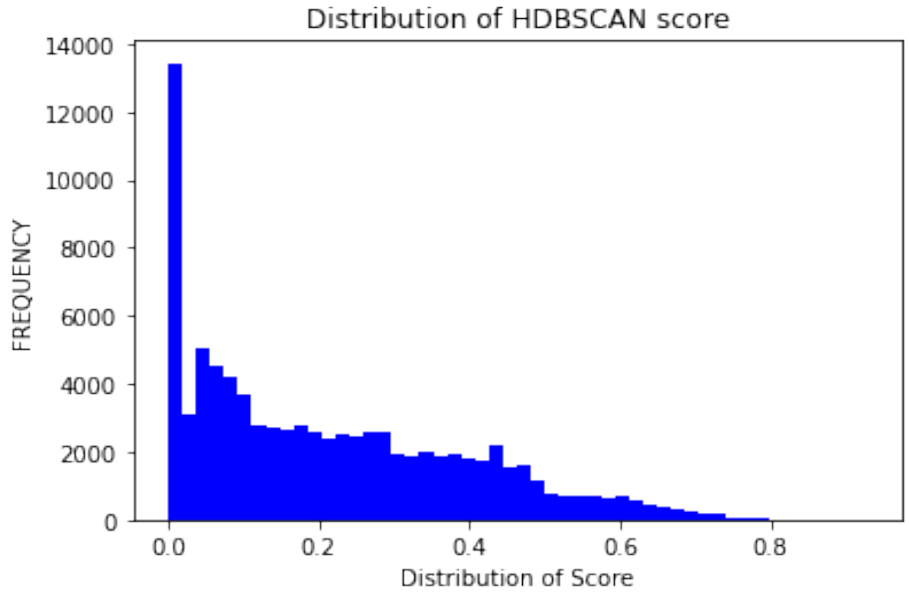


We have trained after splitting the data points in train and test of ratio 80:20 and predict it on test data, we got 0.712 score value and mean absolute error is 0.68 after removal outliers.

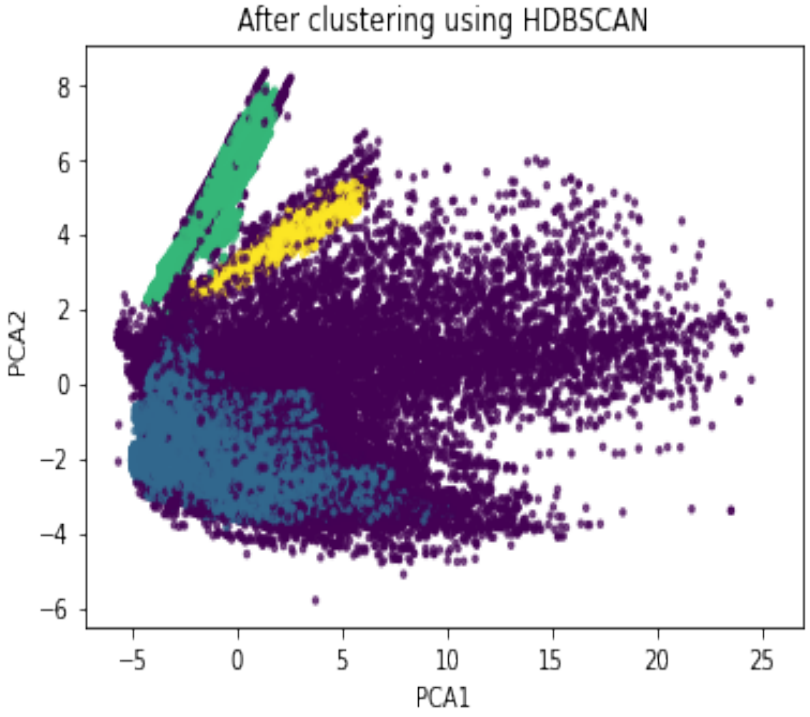
HDBSCAN Clustering (Hierarchical Density-Based Spatial Clustering of Applications with Noise).

- This clustering method is also one of the best method, It detect arbitrarily shaped cluster. It arranges the data points into hierarchies of cluster within clusters, It uses density of neighboring points to construct clusters and allowing noise points to be identified and excluded from clusters.
- It gives more core clusters based on the value of hyper parameters. We have used this method because it gives the expected clustering, however the other method such as K means performs poorly and fails to group the data into clusters and also we don't have ground truth to analyze the clusters so in that case HDBSCAN performs best.
- After finding the cluster using HDBSCAN, we use cluster anomaly score to find the anomaly. More the values, the chance of anomaly is higher, on other hand, lesser the score the probability of being an anomaly is less.

- The below plotted graph shows the distribution of score for anomaly.

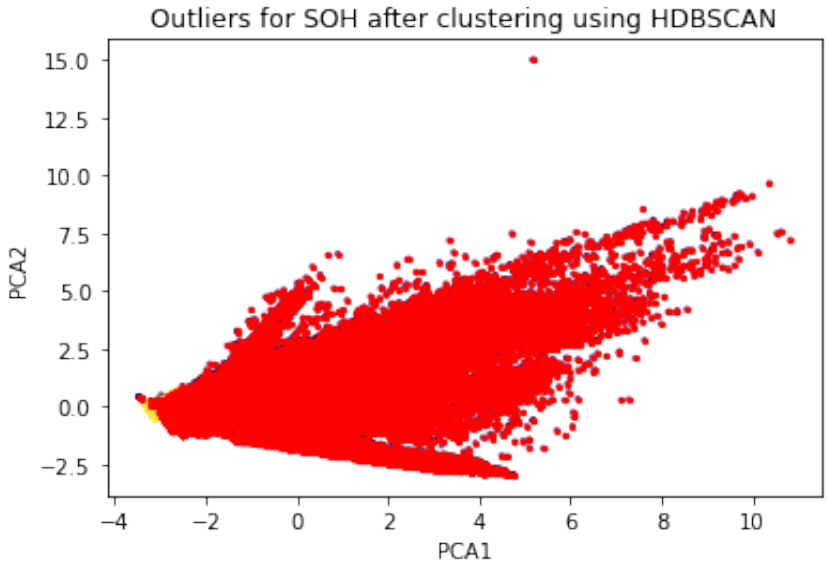


Based on the IQR anomaly detection method, we set the threshold score in a such way that more than 85 quantile, it will consider as a outlier. So after finding, we got 12810 outliers from total 86192.

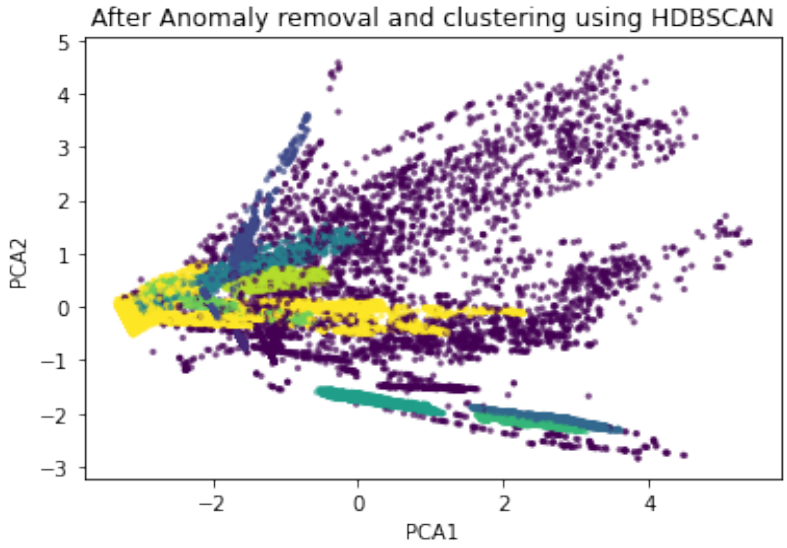


Prediction of Battery Health using HDBSCAN

For the prediction of state of health of battery using HDBSCAN clustering method, we found the anomaly after finding the HDBSCAN score, so if the score is high, it will give more anomaly. Here the threshold is set to quantile of 20 percentage, so all the data points above threshold limit are outliers. The below figure shows the outlier after clustering for state of health of battery.



Now, after removing the outlier and cluster the data points using HDBSCAN as below



We trained the model before and after removing the outliers using different models and found that Random Forest gave better result then other and got score 0.81 and mean absolute error 0.93 before removal of outliers and score 0.79 and mean absolute error 0.88 after removal of outliers.

Result: We have predicted the state of battery health after removal of outliers and found that, Isolation forest performed better than the HDBSCAN.

Conclusion

In this project, we have analyzed the dataset, found the problems and solved using data miningand machine learning techniques. In the supervised machine learning algorithm, for the first problem, which is related to prediction of Operation time. In this problem Support Vector Regression model performed better than the other model with and without PCA on normalized and without normalized dataset.

In Second problem, which is related to prediction of multiclass classification feature 'BatterySupplier', Support Vector Machine gave better accuracy than Logistic regression with and without PCA and SMOTE method.

In un-supervised machine learning problem that is "Anomaly detection of State of Health of Battery and evaluation with different model on prediction of battery health", we have used HDBSCAN and ISOLATION FOREST Clustering methods for anomaly detection. Also, we have compared the result after removal outliers and found that Isolation Forest performed better than the HDBSCAN methods.

Learning from the project

It was good learning while doing this project, we have learnt many concepts while solving theproblems and way to handle challenges. The interesting part was data preparation, while data cleaning and managing to handle missing values, balancing the data, feature extractions werechallenging as well as interesting to find the way of handling these challenges.

We would like to share in brief one of the challenging part, moreover, we have detailed those in the methodology section:

Since the dataset had lots of incorrect and missing values, there were many duplicate samples inraw data. We have removed those duplicate values after understanding the features and relevantdata. Data preprocessing part was most challenging and learning we had found during the project implementation.

References:

1. https://www.kaggle.com/c/data-mining-project
2. https://towardsdatascience.com/speed-up-your-algorithms-part-2-numba-293e554c5cc1
3. https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02231/full
4. https://www.guru99.com/data-mining-tutorial.html
5. https://medium.datadriveninvestor.com/outlier-detection-with-k-means-clustering-in-python-ee3ac1826fb0
6. https://heartbeat.fritz.ai/isolation-forest-algorithm-for-anomaly-detection-2a4abd347a5
7. https://medium.com/learningdatascience/anomaly-detection-techniques-in-python-50f650c75aaf
8. https://petuum.medium.com/scalable-clustering-for-exploratory-data-analysis-60b27ea0fb06