Name: T.P. Dhilip Kumar

Email address: dhilipvasanth@gmail.com

Contact number: 7871745242

Anydesk address: 372 575 103

Years of Work Experience: Nil(Fresher)

Date: 29th Jul 2020

Self Case Study -1: Mercedes Benz-Greener Manufacturing

"After you have completed the document, please submit it in the classroom in the pdf format."

Please check this video before you get started:
https://www.youtube.com/watch?time_continue=1&v=LBGU1_JO3kg

1.0 Project Overview:

1. In an automobile industry,there is a testing phase for every vehicle that comes out from production manufacturing. Safety and reliable testing is a crucial part in the automobile manufacturing process.
2. The Mercedes -Benz automobile industry every day manufactures a huge rate of producing vehicles and finally sends to the testing phase at the final stage in production. Every possible vehicle combination must undergo a test bench to ensure the vehicle is robust enough to keep passengers safe and withstand in daily use. More tests result in more time spent on the test stand, increasing costs for Mercedes and generating carbon dioxide, a polluting greenhouse gas.

2.0 Objective:

The main objective of this project is to optimize/reduce the testing time process of every production vehicle that comes under the test bench. By this optimization it certainly decreases the Carbon dioxide emission associated with the testing procedure.

3.0 Machine Learning Problem Formulation:

The above problem can be solve using Classical Machine Learning techniques to predict the time(target variable) that car will spend on the test bench based on the vehicle configuration.(Independent Features)

1. The type of problem is a supervised learning problem and the model can learn from the labelled data.
2. It is an example of a Machine Learning Regression task thus it should predict the result in a continuous target variable.(time duration of test bench).

3.1 Performance Metrics:

Since it is a regression problem a key performance metric can be used as

1. R^2(coefficient of determination).

3.2 Data source:

Mercedes Benz posted the above problem as csv data format in kaggle platform.

Link :https://www.kaggle.com/c/mercedes-benz-greener-manufacturing

3.3 Data Overview:

The data provided in two csv file formats such as train.csv and test.csv.

1. There are about 386 total features and represented as x0,x1,x2....x386 and each features are anonymized meaning as not physically represented like configuration options such as suspension setting, adaptive cruise control, all-wheel drive,together define a car model.
2. The total features is consist of two format
    a. Categorical variable (8 categorical features)
    b. Numerical binary variable.
3. The product ID columns (unique configuration of vehicle ID).

Research-Papers/Solutions/Architectures/Kernels

1. https://aviationbenefits.org/environmental-efficiency/greener-manufacturing/

   https://www.researchgate.net/publication/235262468_Can_companies_profit_from_greener_manufacturing

   The link refers to the domain knowledge of greener manufacturing.

   General thing about what is greener manufacturing happening in the automobile industry how they are getting profit from this. Here the link explains about the paper how the Greener manufacturing profits in the industry. I will explain the short summary of the content.

   Short summary:

   The automobile companies in European countries to investigate the manufacturing product are environmentally friendly and check whether it is green manufacturing.The specifically giving attention to

industries utilization - like design / methodology approach.The above research was a survey of annual and sustainability reports published by study showed now significant relationship between greener manufacturing and corporate performance, however the trend in decreasing resource, specifically electricity and reducing CO2 was found. Thus this paper aims to quantify the growing trends of environmentally conscious productivity in European companies.

2. “Mercedes-Benz Greener Manufacturing Overview”, Kaggle.com, 2017. Available:https://www.kaggle.com/c/mercedes-benz-greener-manufacturing.
3. https://medium.com/@williamkoehrsen/capstone-project-mercedes-benz-greener-manufacturing-competition-4798153e2476

About the Williamkoehrsen’s problem approach:

Here the above problem he solved his own design modelling he build with new approach

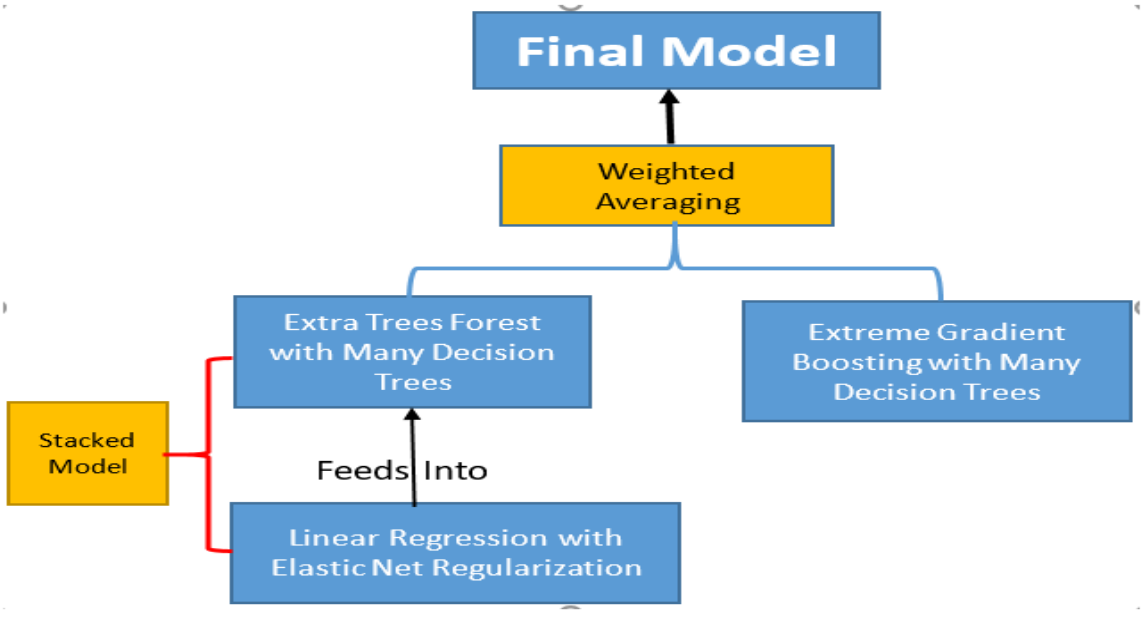


Figure 7: Architecture of Complete Model

The model he approaches is a combination of stacking model and xgboost model. The stack model consists of linear regression with elastic net regularization and extra tree forest with many trees.

On other hand, an ensemble method called Extreme Gradient Boosting. At the top level the final model takes an average of the prediction from each intermediate model.

During training, the preprocessed data will be passed to both intermediate models. The Extreme Gradient model will learn the thresholds for each leaf in the forest of decision trees it trains. The stacked model will first pass the training data through the linear regression, where the model will learn the parameters (weighting) to apply to each feature, then the linear regression will make a prediction for each training point and pass that on as input to the Extra Trees Forest along with the known target values. The Extra Trees regressor will likewise form its own forest of decision trees with the thresholds for each split determined during training. When testing, each new instance will be passed to both intermediate models. In the case of the stacked model, a prediction will be made by the linear regression and then the Extra Trees Regressor will make a prediction based on the output from the linear regression. The Gradient Boosting model will generate its own prediction. The overall prediction will then be an average from the two intermediate models.

By this model implementation the final result he got is

	Stacked Model	XGBoost Model	Final Model
R^2 on training data	0.59182	0.58349	0.58965
RMSE on training data	8.2354	10.9874	9.6785
Median Prediction (s) on train data	100.998	100.964	100.974
R^2 on test data	0.50106	0.53445	0.56705
Median Prediction (s) on test data	101.219	101.131	101.155

4. https://www.kaggle.com/sudalairajkumar/simple-exploration-notebook-mercedes

   The link explains about the exploration of the data different visualization applied for each feature involved in the problem.

5. https://github.com/himanshuknegi/mercedes_benz_greener_manufacturing

   Himanshu Negi's how he approaches the problem is in three approach: In the features engineering part

   First approach:

   There are about 386 total features he applied some dimensionality approach like

   a. Truncated SVD
   b. Principal component Analysis
   c. Fast independent component analysis
   d. Gaussian Random projection
   e. Sparse Random Projection

Here I have learned the above new types of dimensionality reduction techniques that he applied in the features in the above project.

MODEL	R^2	R^2(KAGGLE_SCORE_PUBLIC)	R^2(KAGGLE_SCORE_PRIVATE)
Lasso Regression	0.58441	0.51797	0.52655
Decision Tree Regression	0.599141	0.53522	0.54524
XGB Regression	0.608679	0.5457	0.55275

From the first approach he got the above result

Second Approach:

In this approach he had done with some following step,

a. He had first removed the outlier data points.
b. He calculated the variance of the features and set a threshold of 0.01 if it exist lower by removing those features. By applying this step he removed 146 features and created one new dataset.
c. In the above made important features he applied correlation with the target variable.
d. He also removed some duplicate features which are showing correlation of approx is 1.
e. The above methods he applied correlation techniques,variation techniques to remove and reduce those features.
f. After that he applied some feature engineering techniques, interaction between the features. Like two way interaction and three way interaction.
g. Finally all the above steps are his second approach and applied to the model.

Third approach :

In the third approach he combined both the above two approaches and combined both the approaches features and made one total dataset and trained the model.

By himanshu's techniques I learned some feature engineering parts.

6. https://github.com/level14taken/mercedes-benz-greener-manufacturing

Here how the manoj's approach the problem,

According to his feature engineering,he used T-SVD and selectkbest features reduction technique. I learned a new technique of hyper-parameter tuning what he approaches the Bayesian optimization in hyperparameter tuning.

Here the link which i learned about the Bayesian optimization technique

https://www.sciencedirect.com/science/article/pii/S1674862X19300047

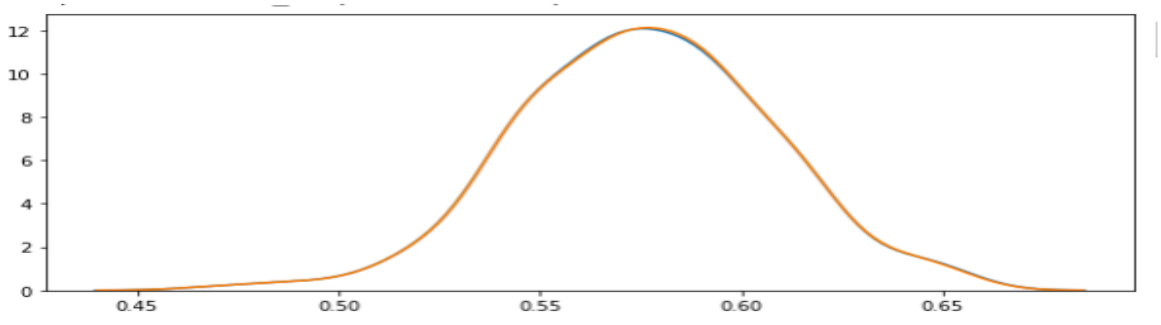
He made three dataset by removing the the duplicate and <0.01 variance features

1. One-Hot Encoded Categorical 2. Label Encoded Categorical 3. Target Encoded Categorical

He put forward that many of the kaggle kernels use the Id features as importance one he performed a T-test to find out if it is true or not.

To do this he took a cross validation score of 250 folds(5 folds,50 repeat) to get the score of the model he trained using the ID and without ID score.

The plot he got of cv score with Id features and without id features.



Distplot between CV Scores of data with ID feature(orange) and without ID feature(blue)

The plot shows that the scores follow a gaussian distribution. He applied scipy.stats.ttest_rel module feed the input as above model and calculated the p values and the alpha values taken as 0.05 if the p values is less than <0.05 thus the accepting the hypothesis thus the Id features is really significant else not.

And he attained the p values as 0.074 which is less than 0.05 alpha and thus the Id features is acceptable. By the way of his conclusion.

Here I learned whether the experiments of Id features are important or not.

7.http://rstudio-pubs-static.s3.amazonaws.com/491189_05618809272041428 0369151c791cf9e.html

The above solution i learned about the different types of data visualization

The author plotted the volcano visualization on both categorical variable and binary variable. And also I can see the box plot advances technique visualizing the data points in the boxplots. This is able to see the boxplot of data points of quartile range as well as the data points displayed in the cartesian space.

8.https://en.wikipedia.org/wiki/Coefficient_of_determination#:~:text=In%20regression%2C%20the%20R2,approximate%20the%20real%20data%20points.&text=Values%20of%20R2%20outside,worse%20than%20a%20horizontal%20hyperplane.

The performance metric that we use in the above project R^2 coefficient of determination I learned about the basic intuition and formula derivation and how it works on the regression model in wikipedia.

9.https://www.linkedin.com/pulse/classification-vs-regression-supervised-learning-dhilip-kumar/

This is my own blog posted in the linkedin discussing the basic regression model intuition which deals with the above problem.

10.https://www.kaggle.com/yohanb/t-sne-visual-3-clusters

The above link is the one of the kaggle notebook in which he made Tsne visualization of total features on train data and test data in that he applies many kind of clustering technique like,

a. Mini batch K-mean
b. Agglomerative clustering
c. Birch clustering
d. Gaussian Mixture

Here the visualization of Tsne of train features and test features of the data.

By the above visualization he made up three clusters and and applied various types of clustering techniques what i mentioned above

First Cut Approach

Step by step approach for the problem:

1. First way to know about the data is to do some preprocessing technique whether the dataset contains the NULL values,duplicate categories.
2. To perform some Exploratory Data Analysis about the features and visualizing how the features are distributed
    a. Univariate analysis plot
    b. Bivariate analysis plot

Plotting some Box plot, violin plot on the categorical variable check whether the categories overlap or well seperate with each other.

Plotting the distribution of the target variable and check how well it is distributed and also visualizing the outlier values.

3. Performing the feature Engineering part:

a. The Categorical variables are converted into one hot encode representation and also here i am going to try with response coding what i learned in course assignment.
b. Perform Tsne to visualize the data points and also apply the pca dimensionality reduction to check how well the model performs.

Other than this any type of features engineering part mentioned by a mentor can also apply to that to improve the performance of the model.

4. Machine Learning Model implementation:

Try with different type of Regression model algorithm and apply the hyperparameter tuning to select the best parameter trade off values to avoid the overfitting and underfitting of the train and test data.

These steps are my basic initial idea before approaching this problem and to add while working on the problem some more idea will be added once the mentor suggests to me to perform a good score.

Notes when you build your final notebook:

1. You should not train any model either it can be a ML model or DL model or Countvectorizer or even simple StandardScalar
2. You should not read train data files
3. The function1 takes only one argument "X" (a single data points i.e 1*d feature) and the inside the function you will preprocess data point similar to the process you did while you featurize your train data
    a. Ex: consider you are doing taxi demand prediction case study (problem definition: given a time and location predict the number of pickups that can happen)
    b. so in your final notebook, you need to pass only those two values
    c. def final(X):
        preprocess data i.e data cleaning, filling missing values etc
        compute features based on this X
        use pre trained model
        return predicted outputs
    final([time, location])
    d. in the instructions, we have mentioned two functions one with original values and one without it
    e. final([time, location]) # in this function you need to return the predictions, no need to compute the metric
    f. final(set of [time, location] values, corresponding Y values) # when you pass the Y values, we can compute the error metric(Y, y_predict)

4. After you have preprocessed the data point you will featurize it, with the help of trained vectorizers or methods you have followed for your train data
5. Assume this function is like you are productionizing the best model you have built, you need to measure the time for predicting and report the time. Make sure you keep the time as low as possible
6. Check this live session: https://www.appliedaicourse.com/lecture/11/applied-machine-learning-online-course/4148/hands-on-live-session-deploy-an-ml-model-using-apis-on-aws/5/module-5-feature-engineering-productionization-and-deployment-of-ml-models