Predictive Vehicle Maintenance

Jaskaran Singh, Karthik Vasireddy, Akhil Masa, Kaushik Junnuri, Mohit Mathur

University at Buffalo, School of Engineering and Applied Science

Abstract— **In some industries, including the automotive industry, Predictive maintenance is becoming more and more important. Especially because the focus is shifting from product to service-oriented operation. It requires to be able to provide uptime guarantees to the customers. Diagnostic and predictive maintenance methods require extensive experimentation and modeling during their development. This is not possible if the complete vehicle is considered as it would require a lot of engineering resources. This thesis provides the results of predicting the forthcoming values of most significant attributes that lead to failure in engine by using time series OBD2 data.**

In this paper, we discuss the concept of trouble codes we learnt via ODB data. We developed model of predicting the status of the features that play key role in engine breakdown. Predicting the status of these components, can help in providing timely car maintenance reminders to vehicle owners and significantly reduce the chances of accident that might occur due to serious mechanical failure.

Keywords—vehicle, maintenance, vector auto regression, OBDII, Random forest, engine coolant

I. INTRODUCTION

It is also imperative to ensure high transport efficiency in order to mitigate low margins and a high turnover in today's traffic. An unexpected increase in fuel prices, vehicle failures etc. can easily turn a profit into a loss. Carriage companies can increase their competitiveness and remain profitable by continually monitoring their transport efficiency which can be enabled by Fleet Management Software's (FMS) provide effective haulage management via advanced Intelligent Transport Systems (ITS).

Due to the increased use of FMS systems, haulers are becoming increasingly concerned about vehicle reliability and therefore availability. A continuous improvement in reliability will be required, and the fierce competition in haulage will drive the demand for less unplanned stops, as most of their other operations are already optimized. In addition to vehicle quality and maintenance actions, driver training can also influence vehicle reliability. Although it can reduce the probability of unplanned stops, preventive maintenance may increase maintenance expenditures. Lease programs or service contracts enable haulers to be more predictable when it comes to vehicle expenses through a relationship-based business model. Maintaining vehicles is likely to improve with the vehicle manufacturer taking over maintenance responsibilities. Other customers have experienced failures and maintenance strategies based on their expert knowledge of the vehicle. When it comes to experience and expertise, this information places manufacturers above even the largest haulers. Manufacturers are still challenged by relationship-based business models. Profit will be large if more vehicles and parts are sold. A relationship- based business is in backwards comparing to this. The profit will be large if less parts are sold. Our motivation is reducing the overall maintenance cost by getting timely remainders, improve life span, optimize spare part inventory, minimize warranty cost.

The National Motor Vehicle Crash Causation Survey conducted by NHTSA (National Highway Traffic and Safety Administration) found that 44,000 motor vehicle crashes surveyed, were due to mechanical failures and other vehicle related problems. Hence, we aim to minimize crashes due to vehicle failure as well by notifying car owner in advance to take his car for maintenance before components unprecedently breaks down.

II. RELATED WORK

There have been several papers written in the field of fault and anomaly detection in automotive systems and ensemble-based anomaly detection.

The paper [V. Venkatasubramanian, R. Rengaswamy, K. Yin and S.N. Kavuri, A review of process fault detection and diagnosis: part I: quantitative model-based methods, Comput. Chem. Eng., 27 (3) (2003), pp. 293-311] focuses on general survey of fault detection. The detection of faults or anomalies in vehicles has been addressed in several publications.

In the paper [F. Cong, H. Hautakangas, J. Nieminen, O. Mazhelis, M. Perttunen, J. Riekki and T. Ristaniemi, Applying wavelet packet decomposition and one-class support vector machine on vehicle acceleration traces for road anomaly detection, Advances in Neural Networks ISNN 2013 (2013)] anomaly detection is used on vehicle data for the purpose of monitoring road condition. Using training set recordings of from drives in normal operation mode, potholes are identified as anomalies. In contrast to our work, this approach focuses on a specific type of anomaly which differs from the detection of faults, where different types of potentially unknown anomaly types can occur.

In the paper [M. Svensson, S. Byttner and T. Rognvaldsson, Self-organizing maps for automatic fault detection in a vehicle cooling system, 4th International IEEE Conference, 3 (2008)] and [S. Byttner, T. Rögnvaldsson and M. Svensson, Consensus self-organized models for fault detection (COSMO), Eng. Appl. Artif. Intell., 24 (5) (2011), pp. 833-839] idea of fault detection for predictive maintenance of commercial vehicles is proposed. Authors compare data from various vehicles are and anomalies are detected in an unsupervised manner in a way vehicle that deviates from others are highlighted. The unsupervised approach (does not incorporate knowledge about the normal or abnormal operation mode) is advocated by the authors of these papers, which varies from the proposed approach in our paper.

III. DATA

Since last few decades Onboard Diagnostic tool (ODB II) has become mandatory in vehicles in USA. It has the capability to log numerous vehicle performance related metrics. Today, with advancements in machine learning and data science we have the power to analyze the historical ODB II data, collected from car trips and process it to derive valuable insights.

A. On Board Diagnostic data

In consideration of their constant movement, large scale data acquisition on vehicles is difficult. Large scale on-board logging solutions are costly to develop and produce, and not suitable for the transport sector. Vehicle on-board data is sent through a CAN network and is used for vehicle control and status signaling. DTC (diagnostic trouble codes), commonly known as OBDII codes, are warning signals generated by your car's computer system. Each system in your vehicle has its own set of parameters. When the car recognizes that the problem exceeds certain limits, a trouble code is generated. These codes are used by technicians to diagnose and fix car problems.

Data was collected from the Kaggle of Mr. Cephas Barreto, which he gathered during his thesis of his master's degree. The data set was made up of data collected from 2 different experiments. First, one being of 14 drivers' On-Board Diagnostics (OBD) data collected while they were driving their automobiles on daily routes. In second experiment, the data was collected from 4 Drivers driving a particular vehicle, in similar way to first experiment. Data was recorded specifically for the Power Train Module and includes the vehicles metadata.

IV. APPROACH

We first did the exploratory analysis of data. On Doing Exploratory data analysis we realized there are certain limitations that will always be there with respect to data being collected by researcher by doing real time experiment via recording various real time Car Drives data on ODB2 device, due to uneven ratio of Trouble codes to the no anomaly in the given dataset. We filtered out the data, in such a way so that we get concrete results in terms of anomaly prediction and tried to minimize any sort of bias that might occur while predicting.

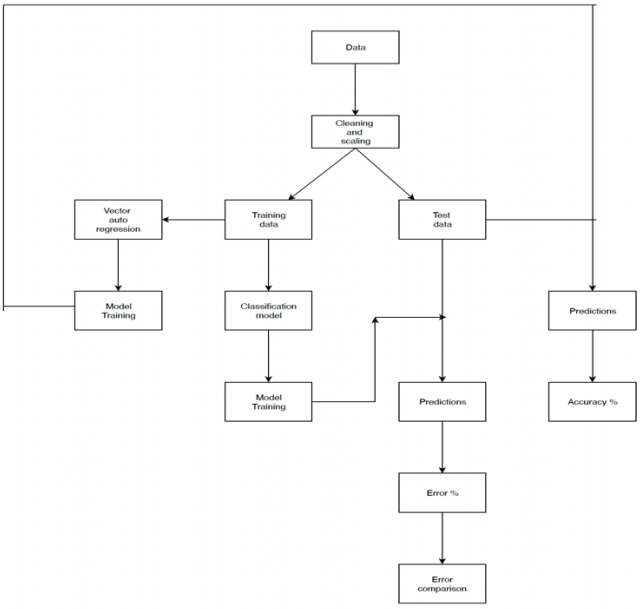

Fig 1: Represents approach of our project

The above flow chart explains our approach logically, following are the expansion of terms used in flow chart:

- Data: This is the raw Data that will be given to further to do cleaning and scaling.
- Cleaning and scaling: Data may contain null values and outliers, to remove those, cleaning is necessary which makes data with no holes in observations and feature scaling is to stabilize the data by scaling all attributes to same range, when any attribute have different range of values compared to one and other.

- Training data: This is pre-processed data which will be used to train the model.
- Test data: This cleaned data which will be used to get results of model after training it.
- Classification model: After data has cleaned, with trained data, different classification models such as random forest, KNN, bagging, boosting, CART, logistic regression and Naïve bayes to predict the trouble code observations.
- Predictions: Predictions of the data are collected here. All the predictions for the different stocks will be saved here.
- Error%: Error is calculated here by comparing the true value and the predicted value.
- Error Comparison: The error rate of classification models is compared with each other.
- Feature selection: After doing cleaning and scaling for the data, there were 20 variables in data. To know which have better correlation with output variable, we have used VarImp method and correlation matrix to obtain important correlated attributes.
- Model Training: Training model is a process in which a machine learning ML algorithm is fed with sufficient training data to learn from, to get precise results on test data.
- Classification models: The classification models used are used to predict trouble codes.
- Vector auto regression: Vector auto regression is a statistical stochastic model, basically uses time series data, used to find out the relationship nature between multiple attributes as they change over the time.

Following is the detailed approach:

- So, Data consists of about 34 attributes namely, Engine coolant temperature, Speed of vehicle, Intake manifold pressure, Engine load, Air intake temperature, Trouble Code being some of the notable columns in terms of importance for our research paper amongst all the features.
- The data we gathered had many N/A values across certain attributes. We dealt with it in following 2 ways:
    - If the entire column data was null: In this scenario, we had no other option but to remove that column from our dataset, as due to limitation of data it could in no way be of any contribution in our research.
    - If there were some missing values for certain features: In that case we imputed the missing values using Multivariate Imputation via Chained Equations (MICE), which is a package of R and it creates a multiple imputations taking care of uncertainty in data.
- We observed ranges of features across data was on different scale. Hence, to normalize the data we performed min-max scaling.
    - For instance, engine rpm was ranging from 714 to 2016 and engine load had range of 12 to 100. Thus, it was important to perform feature scaling to have data across different attributes on same scale.
- Then we performed feature selection techniques to determine the most important attributes to predict our target variable Trouble Code. Here are the techniques we used to select most important attributes:
    - Using the correlation matrix to find the highly correlated features and removing one of the features out of pair of highly correlated features. For instance, in below obtained chart of correlation matrix we removed Throttle_Pos as it is highly correlated with Engine Load and MAF and it doesn't have much correlation with Trouble Code.

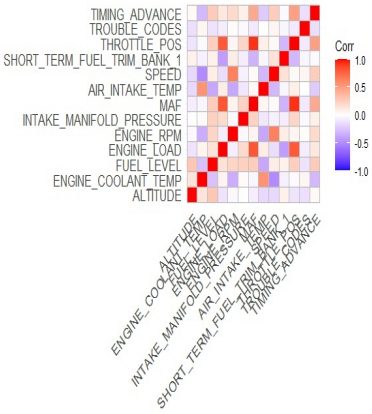

Fig 2: Correlation Matrix

    - Then, to find the features that are most significant in predicting the engine related trouble codes we used random forest method.

- After obtaining important attributes, we have compared the error percentage of various classification models namely, random forest, bagging, boosting, logistic regression, CART, KNN and Naïve bayes to obtain clear cut analysis upon the question of predicting whether observation results trouble code or not.

  Following are the classification models used to predict engine trouble codes:

  - Random forest :It is a supervised learning algorithm, which creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of mean or average.

  - KNN: K-nearest neighbors is used for both regression and classification. It uses data and classifies new data points based upon similarity measures like mean or distance function. The data is assigned to the class which has the nearest neighbors. As you increase the number of nearest neighbors, the value of k, accuracy might increase.

  - Bagging: Bagging or Bootstrap aggregating is it fits base classifiers each on random subsets (which are taken with replacement) of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction.

  - Boosting: It is machine learning method in which, combines a set of weak learners into a strong learner to minimize training errors. a random sample of data is selected, fitted with a model, and then trained sequentially, with each iteration, the weak rules from each individual classifier are combined to form one, strong prediction rule.

  - CART: Classification and Regression tree is decision tree where each node is split in a predictor variable and each node at the node has a prediction for the target variable. It uses cost functions for classification and regression. In both cases the cost functions try to find most homogeneous branches, or branches having groups with similar responses.

  - Naïve bayes: Naive Bayes model are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naïve) independence assumptions between the features.

  - Logistic regression: Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable and nature of target or dependent variable is binary.

    The results of these classification models namely, accuracy, specificity, sensitivity are promising, so we used the best precise result model for prediction for further process.

- Simultaneously, we have developed Vector auto regression to predict forthcoming values of the most significant attributes from time series data. Vector Auto Regression (VAR) was applied in following way.

  The five most significant features that were obtained using random forest were used for VAR. Sensor data is being collected every 4 seconds, so 5 feature columns were converted to time series with frequency of 15 instances per minute. Then, Augmented Dickey Fuller (ADF) and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests were performed to check if the data is stationery. Then, via ADF test we identified first order differencing should be applied to dataset. Below figure illustrates the time series plot of Air Temperature before differencing:

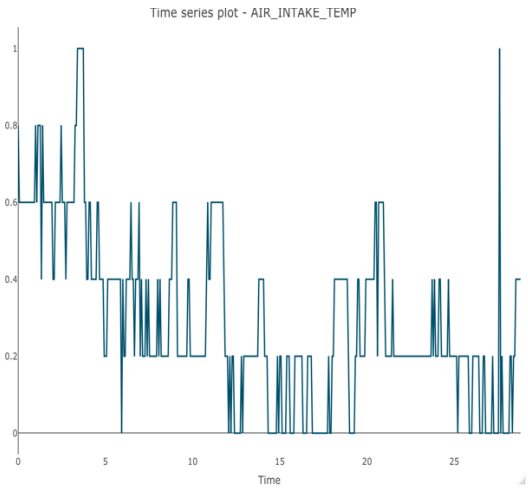

Fig 3: Before Differencing Time series plot- Air Intake Temperature

The figure 3, shows mean and variance are not consistent across various time intervals, which implies trend and/or seasonality. The figure 3, on other hand shows that there are structural breaks in the residuals. Hence, it was needed to perform differencing on given time series data.

After applying the differencing, the following time series plot was obtained for Air Intake Temperature:

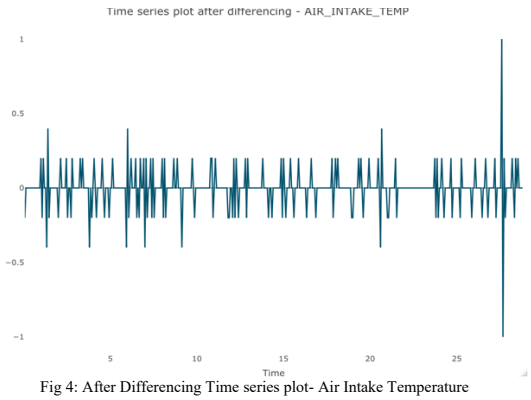

Fig 4: After Differencing Time series plot- Air Intake Temperature

The figure 4, clearly illustrates trend and/or seasonality are removed from data after applying differencing and it is transferred into stationery data.

Then, we identified the optimal number of lag order by using following information criteria's:

- Akaike information criterion (AIC)
- The Hannan–Quinn information criterion (HQC)
- Akaike's Final Prediction Error (FPE) criterion
- Schwarz information criterion (SIC)

For instance, for vehicle id='Control2', lag order came out to be 6.

So, the forecasted state of most significant features obtained via VAR model, will be inputted to classifier to determine whether vehicle engine will need maintenance check-up in future.

V. RESULTS

- Using Random Forest methodology, we found out the features that will be most significant from given dataset in predicting engine related trouble codes i.e., anomalies in engine performance, following are the features we obtained:
  - Engine coolant temperature
  - Engine Load
  - Intake Manifold Pressure
  - Air Intake Temperature
  - Speed

- The performance of various classification models on predicting the engine trouble code was as follows:

	Accuracy	Specificity	Sensitivity
Random Forest	0.97	0.96	0.99
KNN	0.98	0.96	1.00
Logistic regression	0.97	0.95	1.00
Bagging	0.97	0.96	0.99
Boosting	0.98	0.96	1.00
CART	0.98	0.96	1.00
Naïve Bayes	0.97	0.96	0.99

Among the machine learning algorithms we implemented on the dataset knn, boosting and CART perform the best. KNN is often not recomended to use for large and realtime data. Since this application deals with realtime data most of the times, KNN is not suitable to deploy in this case. We have decided to choose boosting, that is Generalized Boosted Regression Modeling because of the flexibility the model provides, we can use several loss functions and provides several hyper parameter tuning options.
The extremely high values that we are getting for accuracy, specificity, sensitivity in our model could be because of the limited data that we had owing to short car trip recording, hence our model might be overfitting on data.

- Using the multivariate time series regression model, we forecasted the future trend of the most significant features responsible for issues in engine, i.e., generating engine related trouble codes. Following are the results we obtained using this methodology:
  - In order to test if there is any serial correlation between error values of the model, we use Portmanteau Test. The null hypothesis of the test states that there is no serial correlation of any order up to lag order p. Portmanteau Test on our model resulted in

p value of 0.2186 which is much greater than 0.05, it implies null hypothesis cannot be rejected.

- Autoregressive Conditional Heteroskedastic-Lagrange multiplier (ARCH-LM) was used to detect heteroscedasticity i.e., checking whether variance of errors is different at various intervals. After performing ARCH-LM test on our model, p value came out as 0.06548 which is greater than 0.05, which absence of heteroscedasticity.

- We plotted the Ordinary Least Square Cumulative Sum (OLS-CUSUM) of the model, to check for structural breaks in residuals. Below figure depicts OLS-CUSUM for the most significant features:

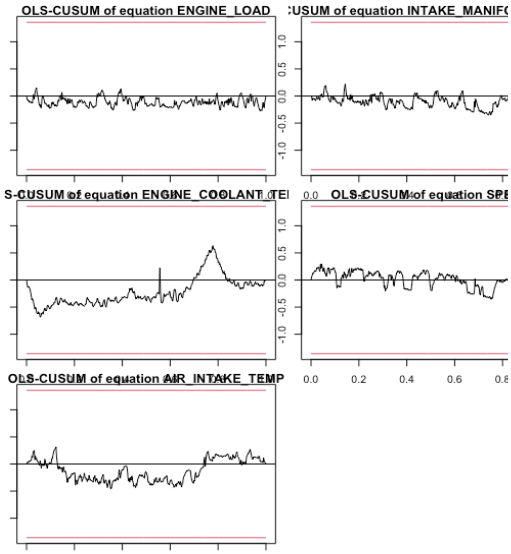

Fig 5: Stability plot obtained after performing OLS-CUSUM

In figure 5, we can see for the plots for all features black line (sum of recursive residuals) never crosses the red line, hence, we can conclude that there are no structural breaks in model.

- Forecast Error Variance Decomposition (FEVD) Analysis, tells the variance in one feature depending on the change in other features. Below is the FEVD plot for our model:

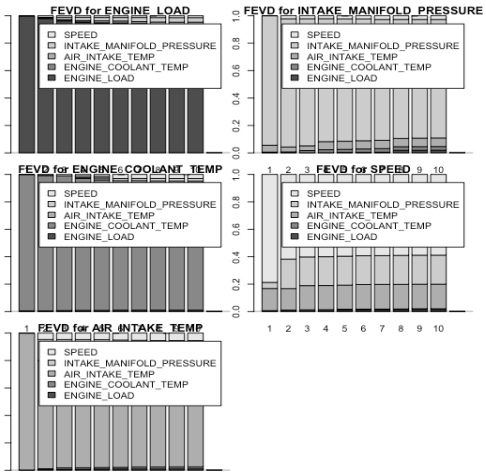

Fig 6: FEVD Plot

From FEVD plot in Fig 6, we can see that, apart from engine load and engine coolant changes in other 3 attributes i.e. Speed, Air intake Temperature and Intake Manifold pressure can be explained by other features.

On applying Grangan-Causality Test on Engine load and engine coolant temperature results in p-value of 0.632 and 0.7434 respectively.

- One of the major drawbacks for our model was due to extremely limited and specific data collected by researcher our model is for a very short trip and dataset is slightly unbalanced as well. So, ideally going forward if we can get the dataset which has recorded trip of far greater duration and is bit balanced as well, we would be able to build a more comprehensive model.

VI. BROADER IMPACT

From the perspective of the manufacturer the concept of vehicle maintenance will lead to a number of advantages that will contribute to the longevity of the vehicle and increased sales.

Some of the factors that can be considered are:

- Life of the vehicle: Vehicle problems can be dealt smoothly if regular maintenance is observed. Repair expenses can be kept to a minimum budget along with reduced wear and tear of the various components of the vehicle. This, in turn, results in extended life of the vehicle and boosts the sales for that make and model.

- Increased Safety: Better safety standards can be incorporated which can always monitor the health and efficiency of the various components of the vehicle. This reduces the risk of sudden vehicle breakdown due to deterioration of the vehicle components. This

feeling of safety assured in the minds of the customers develops trust with the seller and consecutively leads to rise in sales.

- Fluids and Oil Change: This is an important factor considering that decreased friction in components leads to their higher wear and tear and can reduce the performance of the vehicle in the long term. Hence, timely changes of fluids such as coolant, brake fluid, power steering fluid and engine oil prevents the long-term degradation of the overall vehicle performance. Therefore, valuation of the vehicle is influenced by these factors.

- Additional factors such as tire health monitoring and reduced repair costs can be ascertained with vehicle maintenance and makes it an appealing investment to the customer.

From the perspective of the user the better the condition of the vehicle less reliability issues will be faced, and this will enhance the user experience.

Some of the factors to be considered are:

- Increased Safety: With reduced vehicle breakdown and lesser issues with individual components, the user feels more in control of the vehicle and experiences the sense of safety.

- Better car performance: Routine car check-ups lead to increased performance over time and results in lesser expenses if the issues with parts are detected at an early stage.

- Lowers the cost of fuel: Fuel costs are a primary concern for any vehicle user. Regular maintenance of engine components leads to cost effective usage of fuel.

Hence, we can say in nutshell advancements in predictive maintenance techniques, will lead to mutually benefitting partnership for both vehicle manufacturer and owner, and reduce the accidents as well due to mechanical failure apart from potentially increasing vehicle sale for manufacturer owing to better overall vehicle experience for user.

REFERENCES

[1] V. Venkatasubramanian, R. Rengaswamy, K. Yin and S.N. Kavuri, A review of process fault detection and diagnosis: part I: quantitative model-based methods, Comput. Chem. Eng., 27 (3) (2003), pp. 293-311

[2] F. Cong, H. Hautakangas, J. Nieminen, O. Mazhelis, M. Perttunen, J. Riekki and T. Ristaniemi, Applying wavelet packet decomposition and one-class support vector machine on vehicle acceleration traces for road anomaly detection, Advances in Neural Networks ISNN 2013 (2013)

[3] BARRETO, Cephas Alves da Silveira. Uso de técnicas de aprendizado de máquina para identificação de perfis de uso de automóveis baseado em dados automotivos. 2018. 92f. Dissertação (Mestrado Profissional em Engenharia de Software) - Instituto Metrópole Digital, Universidade Federal do Rio Grande do Norte, Natal, 2018

[4] J. Suwatthikul, R. McMurran, R. Jones, In-vehicle network level fault diagnostics using fuzzy inference systems, Appl. Soft Comput. 11 (4) (2011) 3709–3719.

[5] M. Svensson, S. Byttner and T. Rognvaldsson, Self-organizing maps for automatic fault detection in a vehicle cooling system, 4th International IEEE Conference, 3 (2008)

[6] S. Byttner, T. Rögnvaldsson and M. Svensson, Consensus self-organized models for fault detection (COSMO), Eng. Appl. Artif. Intell., 24 (5) (2011), pp. 833-839

[7] A. Theissler, Detecting Anomalies in Multivariate Time Series from Automotive Systems, Brunel University London, 2013 Ph.D. thesis

[8] V. Chandola, “Anomaly detection for symbolic sequences and time series data,” Ph.D. dissertation, Computer Science Department, University of Minnesota, 2009.

[9] C. Marscholik and P. Subke, Road vehicles – Diagnostic communication. Huthig GmbH und Co. KG, 2008.

[10] R. Prytz, S. Nowaczyk, T. S. Rgnvaldsson, and S. Byttner, “Predicting the need for vehicle compressor repairs using maintenance records and logged vehicle data.” Engineering Applications of Artificial Intelligence, vol. 41, pp. 139–150, 2015.