ISBN: 978-981-18-5852-9
WCSE 2022 Spring Event: 2022 9th International Conference on Industrial Engineering and Applications
doi: 10.18178/wcse.2022.04.079

Prediction of Top Key Performance Indicator in Automotive Production System using Data Mining

Akshay Thakur[1], Robert Beck[1], Sanaz Mostaghim [2], Daniel Großmann [3]and Moritz Kuttler[4]

[1]Productivity controlling, Volkswagen AG, Wolfsburg, Germany

[2] Faculty of Computer Science, Otto von Guericke University, Magdeburg, Germany

[3] Faculty of Industrial Engineering & Management, Technische Hochschule - Ingolstadt, Germany

[4] Strategy & Operations, Porsche Consultancy GmbH, Stuttgart, Germany

Abstract. In the past years, the Automotive industry has faced many challenges such as shifting market, digitalization, increased competition, to recent issues of semiconductor shortages. Technology is transforming the automotive industry at a fast pace, and predictive analytics is at its core if harnessed efficiently. Predictive analytics present significant opportunities by using techniques like data mining and machine learning by aiding decision-makers to optimize production processes in the manufacturing plants. Forecasting the critical key performance indicators also ensures a decision support system for the domain experts to take corrective actions. In this research paper, a case study is carried out at the Volkswagen passenger car manufacturing plant in Germany with three years of data to predict one of the top key performance indicators (HPV - Hours Per Vehicle) from the production system. In the global automotive industry, HPV is considered a dominant controlling indicator in production systems. Predicting this HPV ensures robust production planning and provides transparency about the influencing variables so that necessary measures can be taken proactively to improve HPV compared to the planned HPV budget. Comparison between different machine learning algorithms such as Decision tree, Neural network, Linear Regression etc., is done to find accurate machine learning models for key performance indicator prediction based on historical data. Case study results indicate that a Neural network can predict HPV with 2.5% relative error based on historical data. A higher coefficient of determination (R^2=0.8) illustrates that the selected model is stable for prediction. The results show that the machine learning algorithms can be effectively used for forecasting the HPV for the Volkswagen plant and understanding the impact of influencing factors on HPV.

Keywords: Machine Learning, Predictive Analytics, Hours per Vehicle, HPV, Key performance indicator, Data Mining, Neural Network

1. Introduction

The automotive industry, with its increasing volume and variety of data coming during the vehicle product life cycle, gives a rising need for advanced analytics in it. These analytics can generate potential by averting risk beforehand, giving them a cutting edge over their competitors, and promoting rapid growth. Analytics based on available historical data can help the automotive manufacturers carry out different tasks such as performance forecasting, budget allocation, scenario-based production planning and supply chain management etc. The challenge is to find the right artificial intelligence tools and technologies to convert these historical data into insight, which can bolster business success.

Automotive manufacturing companies use a productivity management system to understand the state of their production system/network by tracking the performances of the different processes or activities. These data in the productivity management system is measured in terms of Key Performance Indicators (KPI). KPI helps to understand and improve the manufacturing performance by eliminating wastes from a Lean perspective and achieving companies strategic goals [1]. KPI focuses on different aspects of organizational performance, which are crucial for their ongoing and future success. KPI are measured frequently (24/7, daily, weekly, monthly) depending upon the process requirement [2]. As there is an advancement in manufacturing technologies such as immense use of robotic processes, advanced sensors and process globalization, more and more data are flowing into the production system. The huge amount of historical and real-time KPI data flowing in adds a challenge to business success. Developing advanced analytics

Disclaimer: The results, opinions and conclusions expressed in this paper are not necessarily those of Volkswagen Aktiengesellschaft.

capabilities like machine learning can help manufacturing companies identify patterns and trends that were unseen earlier. Predicting these KPIs helps in business success by taking advantage of future opportunities and reducing risks in the future. The KPI prediction is generally carried out using predictive analytics, i.e., salient methods like probabilistic models, machine learning/data mining techniques, statistical analysis etc., to identify and distinguish patterns from historical data [3].

Predictive analytics is an advanced spectrum of business intelligence technologies compared to other spectrum like reporting, analysis and monitoring. Predictive analytics, being inductive can handle complex data and at the same time deliver high business value as compared to another spectrum (See Figure 1). Predictive analytics is categorized into two types: model-based method and data-based method. Model-based methods use theoretical statistical knowledge such as the specification of the relationship between different variables, model-specific assumptions etc. On the other hand, the data-based predictive modelling method does not rely on any prior models or profuse assumptions. They use different machine learning methods such as Support vector machine, Artificial neural network, Random forest for making data predictions [5]. Data-based learning has a higher capability to make accurate predictions on real-world problems as they benefit from constant learning over time. But these models need to be trained, updated efficiently and adequately [6]. [10], [11], [12], [13], and [14] used data-based modelling in their research. The selection of different machine learning algorithms depends on project-specific requirements, data type, dataset structure, outliers in data, computation power etc. [7].

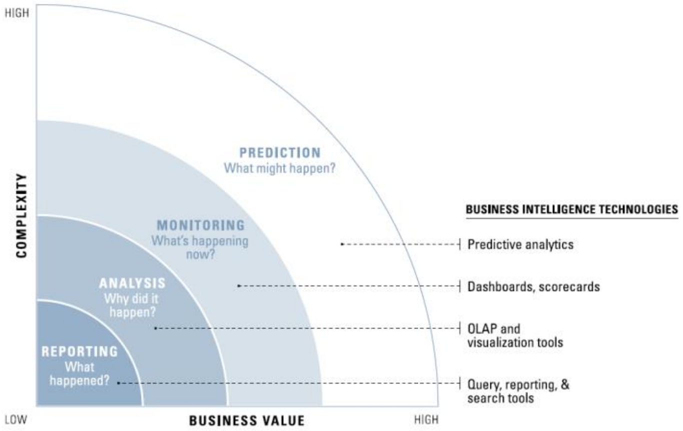

Fig. 1: The Spectrum of Business technologies [4]

This paper is organized as follows: Section 2 presents a literature survey about predictive analytics for KPI, Section 3 is about the methodology used in the case study, Section 4 describes the case study along with the data and experiments carried out for the KPI prediction, Section 5 presents the results and comparison between different machine learning models, and finally, Section 6 summarizes and concludes the paper.

2. Predictive Analytics for KPI

In the predictive analytics field, when it comes to KPI application, it is generally applied in two different ways: KPI selection and KPI prediction. In the KPI selection application type, the goal is to identify suitable and critical KPIs associated with the business objectives from the KPI cluster, so the organizations can focus more on them than other KPIs. KPI prediction application focuses on the prediction of the desired KPI based on historical value. This research focus is limited to the KPI prediction application type.

"Cross Industry-Standard Process for Data mining (CRISP-DM)" is the standard data mining methodology established in 1996. It is used usually for predictive analytics as it is suited for cross-industry due to its design and structure. The process defined for data mining as per CRISP-DM are as below:

- **Project Definition:** In this phase, data mining goals are determined in alignment with business objectives, and then the project plan is formulated.
- **Data understanding:** In this exploratory data analysis phase, data quality is analyzed using different quality check techniques (box plot, heat map etc.). The process is documented along with the dataset description.

- **Data Preparation:** It constitutes data integration from multiple sources, data cleaning and data transformation. Data preparation is a crucial step but time-consuming process. Data selection/exclusion is made based on project goals, technical constraints, or input data types.
- **Modelling:** Models are built, tested and validated based on prepared data. Model validation is done based on accuracy and model generality. In the case of multiple models, models are ranked as per evaluation criteria.
- **Model Evaluation:** Models are evaluated by their performance level dependent on business requirements.
- **Model Deployment & Management:** Models need to be managed and maintained after deployment to sustain/ improve their prediction accuracy. Model management also helps in minimizing redundant activities and standardizing toolsets.

Apart from CRISP-DM, there are similar data mining methodologies like SEMMA (Sample, Explore, Modify, Model & Assess), DMAIC (Define, Measure, Analyze, Improve& Control) etc. As per the 2007 survey (167 respondents) who have implemented predictive analyses, only 15% used CRISP-DM, and the majority were found using their own methodology (52%) [4] [8] [9].

As per the literature survey carried out by Thakur (2020) [9], limited research papers (less than 10) were found in the field of KPI prediction (for the industrial sector) using predictive analytics. The reason for limited research can be associated with companies trying to protect confidential internal KPI data. Oliveira (2015) [10] tested various models such as Random forest, Support Vector Machine (SVM), M5 (modified regression tree algorithm) and partial least square to predict HPV for one of the Volkswagen plants in Europe based on historical data. Prediction accuracy for the model is verified using the Root Mean Square Error (RMSE) value and relative error. The research was carried out with 71 samples and 6 predictors using four months of data. The array of KPI considered was very limited to 2 (Attendance and Volume). This is one of the limited research found where KPI prediction is made in the automotive production systems. The limitation of this research is that it considered limited influencing parameters, as well as model training and testing was carried out on a small dataset. This model cannot predict HPV for a long time duration (more than two weeks), which is overcome by the proposed case study model. Another similar research was found in the educational field, where Gulati (2015) [11] predicted the dropout of students in open courses from a University. In this research, rules-based classification techniques with the help of CRISP-DM were used, such as Jrip, NNge (Nearest neighbor like algorithm), conjunctive rule, DTNB (decision table/Naive Bayes hybrid classifier) and PART (partial decision tree algorithm). 32 initial KPI were decreased to 6 after pre-processing and feature selection. Evaluation of the classification model was done based on three criteria's: (1) by considering the top 10 attributes, (2) considering all attributes, (3) applying data balancing algorithm on selected attributes, and using this data for classification.

Ge (2018) [12] used a different approach of distributed predictive modelling to predict and diagnose KPI in plant-wide processes. This research also highlights that monitoring for plant-wide processes has been a quite trendy research topic in recent years. Still, comparatively less research is done in the prediction and diagnosis of KPI in plant-wide processes. Different processes in manufacturing plants may have a simultaneous and cumulative effect on the final products, making predictive analytics a challenging task. In this research, the complete plant is initially divided into blocks so that KPI data can be extracted efficiently using Principal Component Analysis (PCA) and later used for regression models. Also, the framework consists of the diagnostic part to identify the root cause for KPI performance degradation. The case study was done on Tennessee Eastman Process simulation data. The training and test datasets have 800 and 500 samples, respectively. After the multi-sample rate of input and output variables, they were reduced to 160 and 100.

El-Mongy (2013) [13] used a hybrid method to predict key performance indicators in a balanced scorecard considering the relations between strategic objects. The Balanced Scorecard (BSC) has KPI from four perspectives, i.e., financial, customer, internal business process and development perspective. The Hybrid method proposed in the research consisted of a fusion of association rule and prediction model. Association rule discovers relation between KPI based on past data, and then this rule is fed into Fuzzy Logic Component to predict the KPI values. Side by side, prediction is made using a Neural network, and then this prediction is fused with the output from Fuzzy logic input using a Decision Tree. The case study was done on BSC with four KPIs to predict the fifth KPIs with 532 observations.

Wetzsteina (2011) [14] implemented a data mining approach to showcase KPI dependencies on process and Quality of Service (QoS) metrics. Decision trees were used to identify critical influencing factors for process performance from a total of 31 KPI. The case study determined that the decision tree has the risk of hiding influential factors due to multilevel dependencies between KPI. Another issue with it is that as the tree gets bigger, different influencing factors are included in it, which has an insignificant impact on the main process KPI.

Above mentioned research and case studies indicate that data mining can help in KPI prediction and identifying different influencing parameters so that corrective actions can be taken beforehand for improvement. But all the research has used limited KPI data (less than 1000 samples) in terms of timeframe. Also, hybrid models [13] and distributed predicted modelling [12] are project-specific models and cannot be implemented for all the KPI in general. A worldwide survey was done by Transforming Data with Intelligence (TDWI) in Quarter 2/2018 also indicate that predictive analytics is majorly used for direct marketing (52%), retention analysis (52%) and cross-sell (49%). The survey shows no mention of KPI prediction in the industry, apart from 37% using it for Quality assurance topics and 34% using it for predictive maintenance (total - 244 respondents). The survey also depicted that organizations want to use divergent data for predictive analytics since it adds value to prediction. But lack of machine learning skills and more focus on Business intelligence activities like reporting and dashboards in the organizations prevented them from using predictive analytics [15].

3. Methodology

Below are a few of the regression models used in the following case study:

3.1. Linear Regression

Linear regression is classified into two types: simple regression and multivariate regression. Regression models are used for two purposes: data forecasting and determining the causal relationship between variable and predictor. Simple linear regression is mathematically represented in Equation 1, where only one independent variable is present.

$$y = ß_o + ß_1 x + \varepsilon \tag{1}$$

In the multivariate linear regression technique, more than one independent variables are used to do forecasting. It is represented mathematically in Equation 2, where all the parameters are in the matrix form.

$$y = ß_o + ß_1 x + \cdots . + ß_m x_m + \ \varepsilon \tag{2}$$

Polynomial regression is a special case of Multivariate linear regression, in which the relationship between dependent and independent variable is modelled as an n$^{\text{th}}$ order polynomial in the curvilinear form. Polynomial regression is mathematically represented in Equation 3, where h is the polynomial degree.

$$\mathrm{y} = ß_\mathrm{o} + ß_1 x + ß_2 x^2 + \cdots . + ß_\mathrm{h} x^h + \ \varepsilon \tag{3}$$

The linear regression model has the benefit of being simple to understand and computationally efficient. But their performance can be easily affected by outlier presence [16].

3.2. Neural Network

Also known as Artificial Neural Networks (ANN), they are based on the biological nervous system and working of Neuron. They are used to represent the convoluted relationships between variables by learning, adapting and adjusting the weights between different nodes, known as backpropagation. These nodes (also called units) are arranged in different layers, where the first layer is the input layer, and the last layer is the output layer. The last layer performs aggregation functions of all the weights of the previous layer. The aggregation function at times has a transfer function, which does data scaling. The neural network model uses training data to calculate the error between the predicted value ($\bar{\mathrm{Y}}$) and the actual value (Y). This error (e) is then used to adjust the weights between different units until the error falls within the desired range, as shown in mathematical Equations 4 and 5.

$$e = Y - \bar{Y} \tag{4}$$

$$\mathrm{w} = w^{'} + \lambda * e \tag{5}$$

λ in Equation 5 is the learning rate. The weight adjustment is made by fractional value (λ). New weight (w) after adjustment is the sum of old weight (w′) plus the product of learning rate and error proportion as shown in mathematical Equation 5 [7]. While optimizing the neural network, different parameters can be optimized, such as a hidden layer, training cycle learning rate, momentum, decay etc., to improve prediction accuracy. A Neural network has the limitation that it cannot handle missing data, as well as they are black boxes, i.e., understanding influence of independent variables on a predictor variable is not possible [7].

3.3. Decision Tree/ Regression Tree

The decision trees can be used for regression (also known as regression trees, which are adaptations of decision trees) as well as classification. It breaks the dataset into smaller subsets while the decision tree associated with data is incrementally developed. The decision tree consists of decision nodes and leaf nodes. Nodes represent the splitting rule for one specific feature. For the regression model, it separates them to reduce the prediction error optimally. Nodes are built until the stopping criteria for the decision tree is met. The decision tree size is controlled with the help of different parameters such as information gain, gain ratio, distance-based measure, etc.. Pruning is a common technique used for avoiding overfit of the decision tree with the training data. The most common decision tree algorithm used is ID3. To identify most effective partition/split ID3 uses conditional entropy. ID3 algorithm is described below:

- Step 1: Entropy $H(a_i)$ for each attribute (a_i) is measured, and the smallest entropy is selected.
- Step 2: As per values in attribute (a_i), the dataset is divided, and corresponding sub-nodes are created. Sub-node act as terminal sub-node if all the data belongs to the same class.
- Step 3: In case its non-terminal sub-node, the next attribute is chosen (a_j) with the smallest entropy $H(a_i, a_j)$, and the process is continued. Step 2 is repeated for attribute a_j.

Decision trees have the advantage that nonlinear relationships between features do not affect their performance [17].

4. Case Study - HPV Prediction

The data used in this study is from one of the Volkswagen car manufacturing plant in Germany. The case study is aimed to predict one of the top KPI, i.e. ‘Hours Per Vehicle (HPV)’ of the Volkswagen production system based on the historical data of identified influencing parameters (lower level KPI). HPV is the standard KPI in all vehicle manufacturing plants for personnel productivity analysis and is denoted by the below standard mathematical formula (See Eqn. 6).

$$HPV = \frac{Paid\ attendance\ hours\ in\ certain\ time\ period}{Volume\ produced\ in\ the\ same\ time\ period} \tag{6}$$

Monitoring, analysis and controlling of HPV is crucial in the automotive industry. HPV rankings, benchmarks and comparisons are carried out annually between different vehicle manufacturing brands based on their values. Since they are crucial to companies, their numerical values are well kept secret in the automotive industry. This KPI depicts the success of the respective production plant [18]. Presently, the HPV prediction is made manually using statistics based on the volume of vehicles planned. It is found in research that apart from attendance hours and volume produced, HPV is dependent on more than 50 individual parameters, which are interwoven with each other [18]. This gives rise to the use of advanced predictive analytics to forecast the HPV value since they are capable of handling complex data. The identification of the relationship between different influencing parameters and HPV manually is time-consuming as well as labyrinthine.

The primary objective of the case study was to predict HPV value for a single month (with the prediction for each day) based on the historical data. The secondary objective was to gain insights from the prediction so that it can be used in developing the corrective actions for the influencing parameters so that our HPV remains under the target (lesser the HPV value, better for organization). Predicting the HPV beforehand can help in better production planning and understanding which influencing parameter has how much impact on the total HPV.

Influencing parameters for this project were identified after domain expert interviews internally. Other than HPV value, other influencing KPI like HPV budget, volume per vehicle segment, attendance hours, vehicles produced, First time through rate (FTT) per vehicle segment, Jobs per hour (JPH) etc., were considered. Description for a few of them can be found below:

- **Attendance hours -** It includes all paid hours (except a few special cases) of the complete workforce in the plant. The value-added time (manufacturing time) and non-value added time (breaks, personal allowances), are considered in these working hours. Direct employees (manufacturing personnel) and indirect employees (logisticians, maintenance technicians, managers, administrative staff etc.) are part of those hours. [18]
- **FTT -** It is a measure of production efficiency and quality. It measures how many good units were produced as a percentage of total units produced. It is denoted by the below formula (See Eqn. 7).

$$FTT = \frac{(Total\ units - Defective\ units)}{Total\ units\ produced} \tag{7}$$

- **JPH -** It is the average vehicle produced per hour. This gives an indication of how efficient is the production as per planned plant capacity.

4.1. Data Preparation & Analysis

The data for HPV, along with mentioned and a couple more influencing parameters, are collected for three years (2017-2019) for the complete plant (including body shop, paint shop, assembly shop). Data from the different internal systems are aggregated into one dataset. The data collected has 1095 samples, and each sample represents value for each day. Since the project was started in mid-2020, data of 2020 was not considered due to incompleteness. HPV is considered as the target indicator (predictor) and date as Identification (ID) in the available labelled dataset. Exploratory data analysis for all the features is done in Tableau software to study correlations between them. Initial data with 1095 samples is reduced to 610 after data cleaning. Outlier detection, missing value replacement and constrain verification was done in the data cleaning process. Outlier detection was done to remove samples having irregular HPV values. Analysis for this HPV was done separately to identify the root cause for those particular days. Also, all features were plotted against HPV values and then analyzed to identify outliers. After removing the outliers, HPV data appears to be normally distributed.

4.2. Feature selection

The methodology selected for data mining is CRISP-DM, and the focus is on developing a data-based model for prediction. The software used for building a machine learning model is RapidMiner (Version 9.6). RapidMiner was chosen as the platform due to its modelling capabilities, user-friendliness and also due to its straightforward workflow visualization,which is easy to communicate with other stakeholders.

The original dataset has 137 features, excluding Date and HPV values. The data type of all the features is continuous numerical. The working hour/attendance data is divided into 112 features, each feature representing each department in the plant. From a total of 137 features, 37 are removed due to knowledge redundancy and non-value addition to our predictive model after semi-structured interviews with the domain experts. Features with non-zero variance were also excluded from the dataset.

Correlation between different features was measured using Pearson correlation coefficient (r), which measures the stability of the linear dependence.The correlation coefficient has a range from -1.0 ≤ r ≤ 1.0. A value closer to 1.0 or -1.0 indicates that the features are highly correlated with each other. The Pearson coefficient is calculated as shown in Eqn. 8, where $S_x$and S_y are the standard deviations of random variables x and y, respectively [7].

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})}{N*S_x*S_y} \tag{8}$$

Features with a Pearson coefficient of more than 0.85 were removed from the dataset. Also, by plotting the correlation matrix between features and HPV, strongly correlated features were identified and removed from the dataset. Examples of such are ‘total number of vehicles produced’ and ‘total attendance hour’, which were found to be strongly correlated to HPV, which is evident from HPV mathematical formula (see Eqn. 6).

4.3. Model Selection

The predictive analytics technique is a function fitting method, i.e., fitting data into functions. In standard regression models, dependent variable y is predicted by combining predictor variable X into a function, y = f(X). In this case study, we tested different supervised machine learning models such as Neural network, k-NN (k-Nearest Neighbors), Linear Regression, Support vector machine (SVM), Decision tree, Gradient boosted trees and Random forest based on the data type of features (numerical data).

4.4. Model Evaluation

Different machine learning models are tested by diving data into training and testing sets. Training data is used to train the model, and test data evaluates the model. In the case study, training and validation of the model are done using cross-validation with a simple K-fold method (10 folds). In this method, input data is divided into k subsets of equal size. Of the entire set, a single set is used for testing, and other k-1 subset is used for training the model. The cross-validation is then repeated k times, with each of the k subsets used once for testing. Then the average of all the results is done to create a single estimation. This estimated value depicts whether overfitting is occurring or not, which can be missed when using a regular 80/20 data split (hold out method) [19].

Final model selection is done based on the Root Mean Square Error (RMSE) value, Mean Absolute Percentage Error (MAPE) and Relative error percentage. Moreover, a good fit in the regression model can be verified with squared correlation (R^2,also known as the coefficient of determination), which varies from 0 to 1. Values for R^2closer to 1 represent a good predicting model [7].

5. Numerical Results

5.1. Model accuracy

The prediction accuracy with different parameters such as RMSE value, absolute error, and relative error for all the tested models is showcased in Table 1. The relative error is the average of the absolute deviations of the predictions from the actual value divided by the actual value. All the machine learning models have relative errors between 2.5% to 4.7%. It is clear from the results that the Neural network has higher prediction accuracy (indicating low RMSE value and relative error percentage) as compared to the other models. It has around 2.5% ± 0.6% relative error. Also, the coefficient of determination value for this model is 0.8, indicating that the model is not overfitting the data. It also depicts how much variability in the dependent variable is explained by the independent variables. The least accurate performing model was the Decision tree with 4.7% relative error even after optimization. It also had a lower R^2 value.

Optimization of all the models (hyper-parameter tuning) has shown not much impact on the prediction accuracy (in terms of relative error and RMSE value measurement). For example, the optimized linear regression model with Forward selection had similar RMSE and R^2 values, absolute error increased to 1.7 Hours/vehicle from 1.4 Hours/vehicle and relative error too increased from 3.2% to 3.9%. From the results, it is clear that neural networks can be used to predict accurate HPV values in the future. The training and testing model time was also within the desired range, making it more practical for future application.

Prediction for one-month using test data and the Neural Network model is shown in Figure 2 (see next page). It can be seen that the prediction value is very near to the actual value on most of the days. The data samples where the model could not make correct predictions were analyzed. It was found that on those days, the actual volume had a huge deviation (more than 100 vehicles/day) compared to previous days, due to which the model was unable to capture those trends/patterns.The reason for showing the prediction for this month is that in this particular month, the actual HPV value was a bit distant from the budget HPV value still, the neural Network prediction was near the Actual value, which is commendable.

Table 1: Prediction accuracy of different machine learning models

Machine Learning Model	Prediction Accuracy			
	RMSE Value	Absolute Error	Relative Error	R^2
Linear Regression	2.2 ± 0.9	1.4 ± 0.1	3.2 % ± 0.3 %	0.7
Support Vector Machine	2.3 ± 0.0	1.7 ± 1.6	3.9 % ± 3.3 %	0.6
Gradient Boosted Trees	2.5 ± 0.2	1.9 ± 0.2	4.2 % ± 0.5 %	0.7
Neural Network	1.9 ± 0.5	1.1 ± 0.2	2.5 % ± 0.6 %	0.8
k-NN	2.6 ± 0.3	2.0 ± 0.2	4.5 % ± 0.4 %	0.6
Decision Tree	2.8 ± 0.2	2.1 ± 0.2	4.7 % ± 0.4 %	0.5
Random Forest	2.2 ± 0.2	1.6 ± 0.2	3.6 % ± 0.3 %	0.7

All absolute values are in Hours/vehicle.

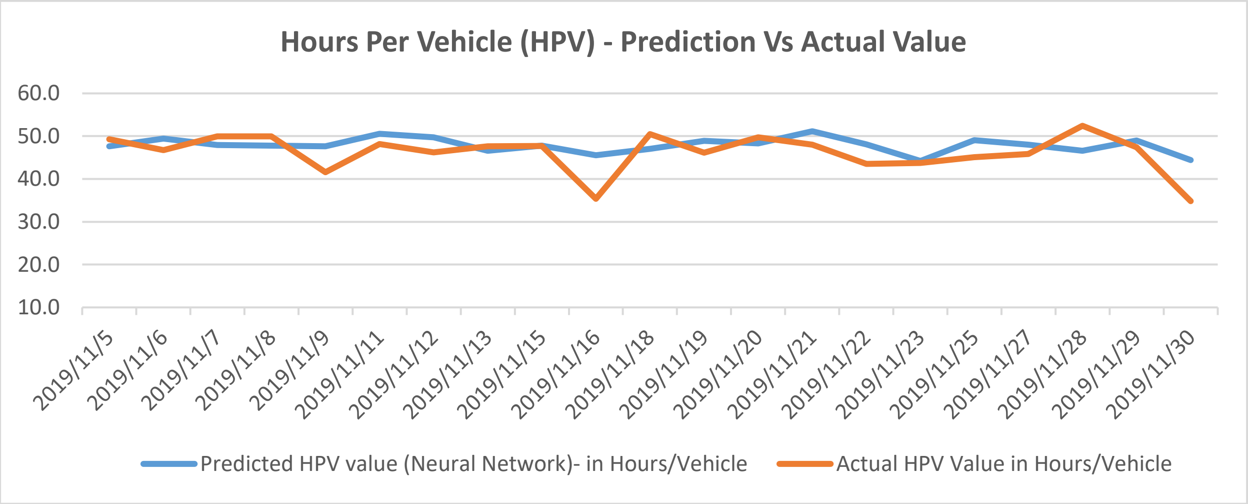

Fig. 2: Prediction value Vs Actual value - Hours Per Vehicle (HPV)

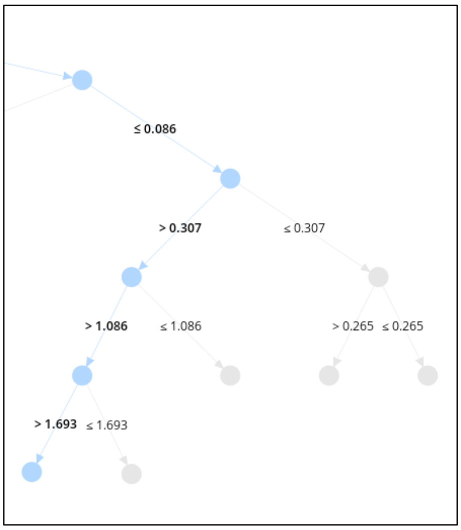

Fig. 3: Branch of the Decision tree depicting influencing factors

5.2. KPI Relationships

Models like Linear Regression and Decision Tree did not provide the best accuracy for prediction but were quite helpful in understanding the influencing parameters for the HPV. From the linear regression model, highly significant influencing factors for HPV can be identified based on the P-value. Influencing factors for HPV having P-value less than alpha value (0.05) are statistically significant for the prediction. Similarly, with Decision Tree's help, HPV value can be narrowed down based on the value range of different influencing factors. It helps in understanding for which influencing parameter, in which range the corrective action needs to be taken, so that desired HPV value is achieved (shown in Figure 3 – see previous page).

5.3. Impact of data

In KPI prediction, one of the prevailing challenges is the availability of abundant data. Since for some of the KPI such as HPV, the data for the influencing parameters cannot be captured by sensors (Ex. attendance hours). Therefore, we have limited data samples (1 sample/day) for training and testing the machine learning model. In Table 2., we can see how the relative error changes if we use only one year of data for prediction instead of three years. The relative error for 2017 and 2018 (Average = 1.2 %) individually are low, but for 2019, it is comparatively high (4.4 %). This can be due to some years having stable production and some years having bit unstable production, such as 2019, where volume per day fluctuates much. But from the below Table 2., it can be seen that with more data use, we can more robustly predict HPV for years having unstable production (such as the year 2019).

Table 2: Effect of data availability on prediction model (Neural Network)

Year of data used	**Prediction Accuracy**			
	RMSE Value	**Absolute Error**	**Relative Error**	**R^2**
2017	0.6 ± 0.2	0.5 ± 0.1	1.0 % ± 0.3 %	1.0
2018	0.8 ± 0.1	0.6 ± 0.1	1.4 % ± 0.3 %	1.0
2019	2.9 ± 1.0	2.0 ± 0.4	4.4 % ± 1.0 %	0.6
2017-2018	0.5 ± 0.1	0.4 ± 0.1	1.0 % ± 0.1 %	1.0
2018-2019	2.5 ± 1.3	1.5 ± 0.4	3.4 % ± 1.0 %	0.8
2017-2019	1.9 ± 0.5	1.1 ± 0.2	2.5 % ± 0.6 %	0.8

All absolute values are in Hours/vehicle.

6. Conclusion

The research paper compares different machine learning models to identify an accurate model to predict the HPV for the next month (with the prediction for each day). The results from the case study clearly show that machine learning models can be successfully used to predict KPI such as HPV in the Automotive production systems. The relative error of all the machine-learning models varied between 2.5% to 4.7%, with Neural Network being the benchmark (2.5%). The Neural network model was also found to be stable with a higher coefficient of determination value ($R^2 = 0.8$). The models like linear regression and decision tree were able to illustrate the relationship between different influencing factors and their impact on the final KPI (HPV). Also, the advantage of the machine learning model being capable of continuous improvement with more data and training is beneficial for future KPI prediction. This predictive analytics application provides a competitive advantage by being proactive in KPI monitoring and controlling rather than being reactive. Predictive Analytics contribute sustainable benefits since it uses data and computational intelligence, which can benefit the manufacturing process in the Automotive industry.

Future work is planned in three steps: (1) To consider more influencing KPI for HPV and more yearly data (2020-2021) for training and testing of the model to compare the model performance and improve it. Also, more options of hyper-parameter tuning are to be explored for prediction improvement. (2) To test this model on other similar Volkswagen car manufacturing plants to check the model's performance. (3) To provide insights learned from this predictive analytics model to build a prescriptive analytics model for HPV. The prescriptive analytics model is planned to be used for optimizing business practices to suit different predicted outcomes.

7. References

[1] ISO/TC 184/SC 5, “Automation systems and integration - Key performance indicators (KPIs) for manufacturing operations management: Part 2 - Definitions and descriptions,” International Organization for Standardization, Standard, Jan. 2014. [Online]. Available:https://www.iso.org/obp/ui/#iso:std:iso:22400:-2:ed-1:v1:en

[2] D. Parmenter, Key Performance Indicators, 3rd ed. New Jersey: John Wiley & Sons, Inc, 2015.

[3] R. Soltanpoor and T. Sellis, “Prescriptive analytics for big data,” Lecture Notes in Computer Science, pp. 245–256, 2016.

[4] W. W. Eckerson, “Predictive Analytics: Extending the value of your data warehousing investment,” TDWI Research, Tech. Rep., 2007. [Online]. Available: www.tdwi.org

[5] Z. Ge, Z. Song, S. X. Ding, and B. Huang, “27_Data Mining and Analytics in the Process Industry: The Role of Machine Learning,” IEEE Access, vol. 5, pp. 20590–20616, Sep. 2017, doi: 10.1109/ACCESS.2017.2756872.

[6] C. Gao, H. Sun, T. Wang, M. Tang, N. I. Bohnen, M. L. Müller, T. Herman, N. Giladi, A. Kalinin, C. Spino, W. Dauer, J. M. Hausdorff, and I. D. Dinov, “Model-based and model-free machine learning techniques for Diagnostic prediction and classification of clinical outcomes in PARKINSON’S DISEASE,” Scientific Reports, vol. 8, no. 1, 2018.

[7] V. Kotu and B. Deshpande, Predictive analytics and data mining: concepts and practice with rapidminer. Morgan Kaufmann, 2014.

[8] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth et al., “Crisp-dm 1.0: Step-by-step data mining guide,” SPSS inc, vol. 9, pp. 1–73, 2000.

[9] A. Thakur, R. Beck, S. Mostaghim, and D. Grosmann, “Survey into predictive key performance indicator analysis from data mining perspective,” 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), 2020.

[10] S. A. P. de Oliveira, “Performance Management Analytics for theAutomotive Industry,” Master’s thesis, Faculdade de Engenharia daUniversidade do Porto, Portugal, 2015.

[11] H. Gulati, “Predictive Analytics Using Data Mining Technique,” in 20152nd International Conference on Computing for Sustainable GlobalDevelopment (INDIACom). New Delhi, India: IEEE, 2015, pp. 713–716.

[12] Z. Ge, “Distributed predictive modeling framework for prediction and diagnosis of key performance index in plant-wide processes,” Journal of Process Control, vol. 65, pp. 107–117, may 2018.

[13] A. M. A. El-mongy, A. E.-d. Hamouda, N. Nounou, and A. W. Abdelmoneim, “Design of prediction system for key performance indicators in balanced scorecard,” International Journal of Computer Applications, vol. 72, no. 8, 2013.

[14] B. Wetzsteina, P. Leitnerb, F. Rosenbergc, S. Dustdarb, and F. Leymann, “Identifying influential factors of business process performance using dependency analysis,” Enterprise Information Systems, vol. 5, no. 1, pp.79–98, feb 2011.

[15] F. Halper, “Practical Predictive Analytics,” TDWI, Tech. Rep., 2018.

[16] D. Maulud and A. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning", Journal of Applied Science and Technology Trends, vol. 1, no. 4, pp. 140-147, 2020. Available: 10.38094/jastt1457.

[17] A. Navada, A. Aamir, S. Patil and B. Sonkamble, "Overview of use of decision tree algorithms in machine learning", in 2011 IEEE Control and System Graduate Research Colloquium, Shah Alam, Malaysia, 2011, pp. 37-42.

[18] M. Weyer, “Hours-per-vehicle controlling – the renaissance of staff productivity,” International Journal of Production Research, vol. 49, no. 11, pp. 3271–3284, 2011.

[19] RapidMiner. GmbH, "Cross Validation – RapidMiner Documentation", Docs.rapidminer.com, 2021. [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/validation/cross_validation.html. [Accessed: 19-Feb- 2021].