IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 07-13

Transport Vehicle Selection Predictor

Shubham Matale[1], Apurva Taru[2], Pratik Mahamuni[3],Prof. Sukhada Bhigarkar[4]
[1](*Department of Computer Engineering MIT College of Engineering Pune, India)*
[2](*Department of Computer Engineering MIT College of Engineering Pune, India)*
[3](*Department of Computer Engineering MIT College of Engineering Pune, India)*
[4](*Department of Computer Engineering MIT College of Engineering Pune, India)*

Abstract: *Vehicle Selection Predictor is a new and important concept in transportation studies. In recent years, prediction model was designed for the prediction of prices of vehicles. In this project, we are trying to build a prediction model for vehicle selection based on its usage. In this project, we consider the problem of vehicle selection for transportation using a multicriteria decision-making approach. The problem includes several conflicting factors which are economic and technological factors. The vehicle has its own specifications/factors like load that it can handle, average of the vehicle, design of vehicle etc. Based on these factors the transport agency should choose a vehicle for a particular delivery. Hence, to predict the selection of vehicle depending on its specifications, the system will be designed by using machine learning algorithms. While designing the model feature selection is a very important aspect. This paper mainly focuses on feature selection and building the predictive model.*

Keywords:*Classification Algorithm, Feature Selection and Evaluation, Machine learning, Predictive model*

I. Introduction

Vehicle selection predictor is an act of predicting the suitable vehicle for the required route for the transport agency. In the transport system, choosing a vehicle for a particular route is a difficult task. As there are many options available while choosing a vehicle, it makes confusing to choose a suitable vehicle.

All the transport management responsibilities, vehicle selection, and allocation are sure to present one of the greatest conundrums. There can be some complicated factors to be consider like the load capability, engine power, fuel capacity of the vehicle, etc. Achieving the goal of getting the best value up front and throughout a vehicles service life while providing the right equipment for the job. By considering this, the system will try to reduce the efforts required for the selection of vehicles with the help of Machine Learning Algorithm.

Transport Management Systems (TMS) are gradually extended with new features to improve reliability, such as planning vehicleroutes. This project will be an improvement over the existing TMS, which will provide the vehicle selection based on machine learning. It will truly automate the transport planning, as the present systems are not effectively able to plan trips of vehicles according to their full efficiency.

Prediction of vehicle depends on various vehicle specifications such as the load capacity of the vehicle, the power of the engine etc.[1]. The system will be designed by using machine learning algorithms. The purpose is to select the vehicle from all other available vehicles for a respective route, by comparative study of popular classification and prediction techniques. This work gives a brief introduction towards techniques, application, and challenges of the prediction system. This project is as important to the individual as much as to the public too.

In this paper, section II overviews previous work done related to transportation engineering and machine learning. Section III discusses problem solution and it gives an idea about overall proposed system's architecture. Section IV discusses about methodologies used. Finally, section V is about conclusion and future work.

II. Relevant Work

In [1], the author explained various factors affecting transportation. The performance, design, and operation of the vehicle are affected by various factors such as a human, vehicle, acceleration characteristic, breaking performance etc. These factors influence the geometric design and design of control facilities. Variant nature of the vehicle and road surface have importance for the smooth and efficient performance as well as the smart planning of the transportation vehicles.

In 2014, R. Prytz, S Nowaczyk, T Rgnvaldsson, S Byttner investigated unsupervised and supervised methods for predicting vehicle maintenance [2]. These methods are data-driven and use big amounts of data which may be streamed, onboard data or historical data from the off-board database. These methods depend on telematics gateway that enables vehicles to communicate with the back-office system.

Author SerhatAydn and CengizKahraman considered the problem of bus selection for public transportation using a hybrid multicriteria decision-making approach [3]. The problem includes various contradictory factors such as social, economic and technical factors.

The author brings integrated approach of fuzzy analytic hierarchy process (AHP) and simplicity of fuzzy VIKOR methodology together. To test the methodology, a case study of Ankara, the capital of turkey is given. In this system, the four-level hierarchy was established along with three experts. These experts are utilized for assessing the pairwise comparison matrices. The weights of criteria were determined using fuzzy AHP and then alternative ranks by fuzzy VIKOR. Sensitivity analysis is also made to see how sensitive decision predicted. It helps to change in parameters of methodology. Buckley’s fuzzy approach is finally implemented to solve the problem.

In 2014 author Visalakshi, S., Radha, V. studied basic concept of feature selection, the process of feature selection, evaluation criteria and various approaches used in feature selection which are applicable to all databases [4]. From the state of art, feature selection methods are flexible and capable of providing a solution to any kind of problem faced when performing feature selection. From the literature study, instead of using filter and wrapper methods separately, better embed both methods for selecting relevant features, improves in classification accuracy, speed up the process and reduce the error rate. Still, there is a scope that genetic algorithm and ensemble (filter and wrapper) are capable of handling multi-dimensional dataset.

In the article [5] author explained how different methods shows different variables are important, or at least the degree of importance changed. The result need not be conflict, because each method has a different approach and perspective for how a variable can be useful.

III. Proposed Solution

The problem of vehicle selection for transportation is solved by using a multi-criteria decision-making approach. The problem includes several conflicting factors which are economic and technological factors.

Fig. 1 shows the basic architecture of the proposed system.It describes modules such as feature selection,separation of training and testing dataset etc. It gives an insight into what are the different modules of the system and how they interact with each other.

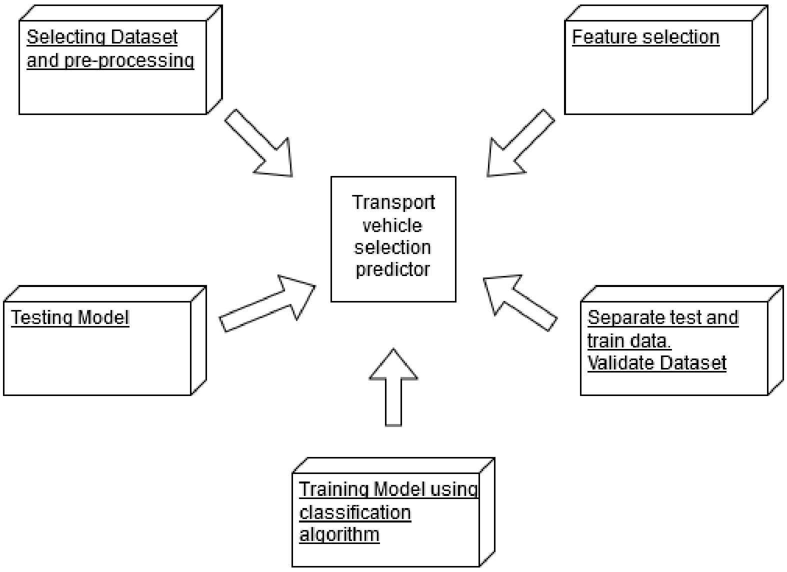
Selecting Dataset and pre-processing
Feature selection
Transport vehicle selection predictor
Testing Model
Separate test and train data. Validate Dataset
Training Model using classification algorithm

Fig. 1. Block Diagram

In accordance with the problem statement, the proposed system will train the model to classify the various vehicles based on various parameters and predict the suitable vehicle for a particular condition. The proposed system will act as a decision support system to the user who needs to select the vehicle. This system will automate the task of selecting the vehicle taking into consideration parameters like location, load type, load capacity etc. As shown in fig. 1, all the modules are explained as follows:

A. Dataset Selection and Preprocessing

In this phase, data of various transportation vehicles will be collected with the following attributes:

- Company Name
- Model Number

- Engine Type
- Engine Cylinder Count
- Displacement
- Max Power
- Max Torque
- Transmission Type
- Gearbox
- Fuel Tank Capacity
- Turning Radius
- Max Speed
- Wheelbase
- Overall Vehicle length, Width, Height
- Ground Clearance
- Payload
- Type of Location

Various data pre-processing methods such as Data Cleaning, Integrating, Transformation, Reduction etc. will be applied to generated data[6].

B. Feature Selection

Feature selection becomes more important when there is a large set of features available. It is not necessary to use every feature for the system. It is necessary to assist algorithm by providing only those features that are really important. Selected feature subsets give better results than a complete set of features for the same algorithm. The methods used for feature selection is wrapper methods which contain two sub-methods.

Wrapper methods: In wrapper methods, a subset of features is used for training the model. Based on the inferences that are drawn from the previous model, it is decided to add or remove features from the subset. The problem is essentially reduced to a search problem. These methods are usually computationally very expensive. Types of Wrapper Methods are:

- Forward Selection: Forward Selection starts with no variable in the model. For all predictors which are not in the model, check their p-value if they are added to the model. Choose the one which affects the result. This method is repeated until no new variables added to the model [4].
- Backward Elimination: Backward elimination starts with all the features and removes the least significant feature at each iteration which improves the performance of the model. This process is repeated until no improvement is observed on the removal of features [4]
- Recursive Feature Elimination: It is a greedy technique which aims to find the optimal solution i.e. best performing feature subset. It continuously creates models and separates the best performing or worst performing features at each iteration. It creates the model with the remaining features until all the features are exhausted. After that, it ranks the features based on the order of elimination of features [4]. Fig. 2 shows, the working of forward selection and backward elimination.

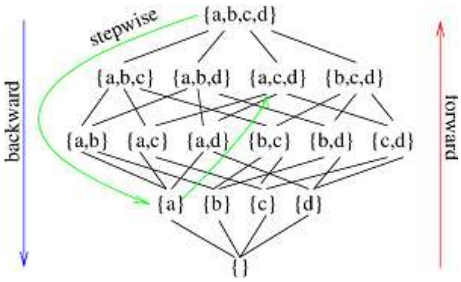

Fig. 2 Feature Selection Methods

For example, if there are 4 features such as [a,b,c,d]. In forward selection initially it start with no feature set. In first step, it adds onefeature and checks for accuracy [a], [b] etc. In next step it adds another feature and check for accuracy with various combination [a,b], [b,c] etc. The combination with most accuracy is selected and another feature is added to check the accuracy [a,b,d], [b,c,d] etc. In similar manner the process is repeated untill the set of best features with good accuracy is achieved [a,b,c,d]. In contrast backward elimination starts with all the features initially and eliminates features one by one according to the accuracy contribution of that particular feature.A step one contains[a,b,c,d]. In further step it contains [a,b,c], [b,c,d] etc. In similar manner steps are repeated to select best features set.

C. Dataset Validation

Validation techniques are used to lower the error rate of the machine learning (ML) model. It is more effective when dataset available is too less. Two proposed methodologies for dataset validation are as follows.

- K-Fold Cross-Validation

In this method data is split into k subsets of equal size by random sampling. In the next step, each subset in turn is used for testing and the remainder for training. The advantage is that all examples are used for both training and testing. The error estimates are averaged to yield an overall error estimate [7]. Error rate is derived from equation 1

$$E = \frac{1}{k} \sum_{i=1}^{k} E(i) \qquad (1)$$

Where, E is error rate.
k in number of subsets.

In fig. 3, the colored part is for testing and remaining is for training purpose. For example, consider initially data is divided in 4 subset. In each experiment one subset is used for testing and remaining for training.

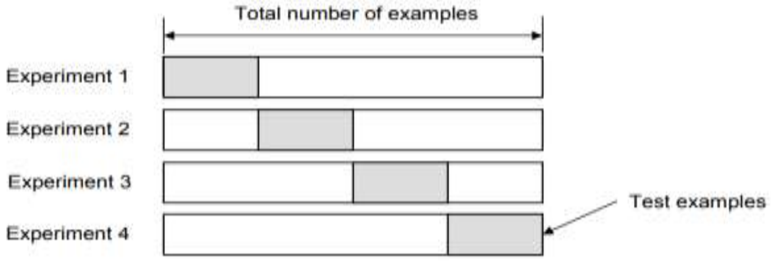

Fig. 3. K-Fold Cross-Validation Method

- Leave-One-Out Cross-Validation (LOOCV)

If the dataset has N examples, then N experiments to be performed for Leave-one-out cross-validation. For every experiment, training uses N-1 for training and one example is use for testing . The average error rate on test examples gives the error[7]. Error rate is derived from equation 2

$$E = \frac{1}{n} \sum_{i=1}^{n} E(i) \qquad (2)$$

Where, E is error rate
N is number of examples.
This method is generally preferred over the K-Fold CrossValidation because it does not suffer from the intensive computation, as number of possible combinations is equal to number of data points in original sample or n.
As shown in fig. 4, a single data point for testing and remaining is used for training purpose. This is repeated for all experiments. Then the error is averaged for all the experiments to give overall efficiency.

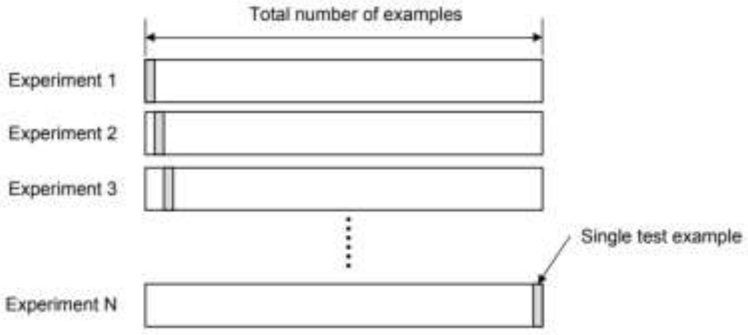

Fig. 4. LOOCV Method

D. Training Model
In machine learning, classification is a supervised learning method. In this method the computer program learns from the data input given to the algorithm and then uses this program to classify new observation.
The first algorithm deals with clustering.

1. K - Nearest Neighbour

K - Nearest Neighbors is used in the field of pattern recognition. It learns by analogy, i.e by comparing a given test tuple with training tuples that are similar to it. The training tuples have n attributes, every tuple represents a point in n-dimensional space. All training tuples are stored in n-dimensional patterns.
K-Nearest Neighbors searches the pattern space for the k training tuples that are closest to the unknown test tuple. Closeness is defined using distance metrics such as Euclidean distance [8].

- Step 1: A positive integer k is specified, along with a new sample. (k is the number of nearest neighbors).
- Step 2: Select the k entries in the database which are closest to the new sample (calculate the distance between those points).
- Step 3: Find the most common classification of these entries (one having the shortest distance).
- Step 4: Assign that classification to the new sample.
  In Fig. 5, the test sample (which is inside the inner dotted circle) should be classified either to the first class of circle or to the second class of triangles.

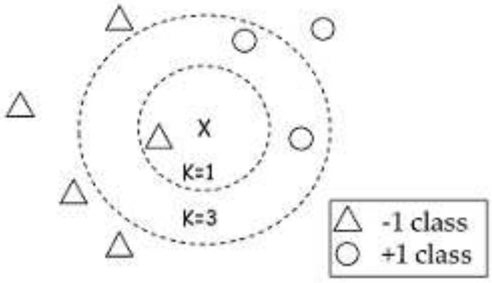

Fig. 5. K Nearest Neighbour

If k = 3 (which is outside the inner dotted circle), it is assigned to the second class because there are 2 circles and only 1 triangle insidethe inner dotted circle. For example, if k = 5 then it is assigned to the first class (1 circles versus 4 triangles outside the outer dotted circle).

2. Neural Network

Neural network consists of many simple processing units, which are wired together in a complex communication network as a human brain. Each of the processing unit is called a neuron. It consists of input signals, weights assigned to each of the input, processing function f which computes the summation of weighted input and output signals [9] .The basic structure of a neuron is as shown in fig 6.

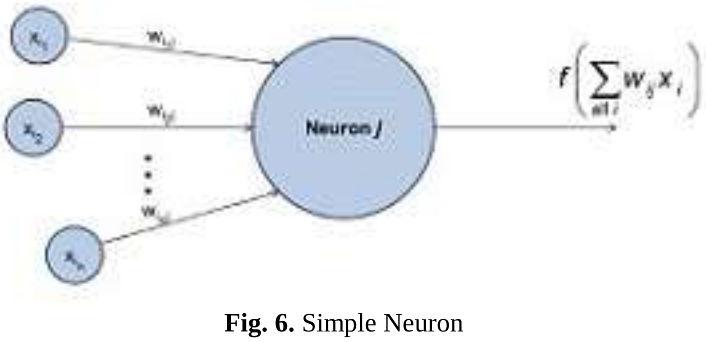

Fig. 6. Simple Neuron

The steps involved in the algorithm are:

- The n inputs are denoted by x(1),x(2)...x(n).Weights are denoted as w(1),w(2)...w(n).
- The input values are multiplied by their weights and those are summed as equation 3

$$E = \sum_{i=1}^{n} w(i)x(i) \qquad (3)$$

- The output is some function y = f(v) of weighted sum.

IV. Discussion

There are overall 17 features of dataset but some of them are not necessary so wrapper methods are applied to select the features. During the experiment, among the 3 wrapper methods, backward elimination has given the better feature set where label is Location Type. Intracity, Intercity, Hilly, Highway are four classes of location type. Intracity represent the particular vehicle is best for within a city or minimum distance transportation, similarly Intercity for two cities. Hilly shows the mountain roads transportation, and Highway is for long distance smooth roads.

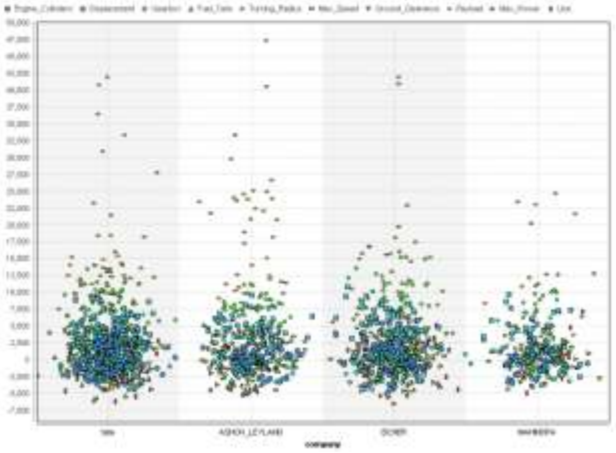

Fig. 7. Scatter Diagram of Selected Feature Set

Fig. 7 shows the scatter diagram of selected feature set. Scatter diagram represent the distribution of feature dataset along the x-axis which represent the company name and y-axis represent the values of features.The selected attributes are as follows:

- Engine Cylinder Count
- Displacement
- Gearbox
- Fuel Tank Capacity
- Turning Radius
- Max Speed
- Ground Clearance
- Payload
- Max Power
- Use

Validation of dataset is important aspect for good accuracy of model. Among the LOOCV and k-fold validation, K-fold cross validation is exhaustive in the sense that it needs to train and validate the model for all potential combination and for moderately large number of subsets, which becomes computationally infeasible. In contrast, number of possible combinations in LOOCV is equal to number of data points in original sample hence it does not suffer from the intensive computation so LOOCV has better performance. It also leads to higher variance and get estimates of test error with lower bias. Therefore LOOCV is to be preferred over K-Fold Cross Validation for validation.

In this paper, two classification algorithms are used; the K-Nearest Neighbor(KNN) and Neural Network(NN). Cross validation is used to classify and validate the dataset. The accuracy output of the both classifieris67.49%and72.71% respectively. KNN iscomputationally expensive as it need to store all the training data, whereas in neural network once model is trained, the training data is no longer needed to produce new predictions.

Table 1 represents the performance matrix for the Neural Network classifier.

	Actual				
	Intercity	Intracity	Highway	Hilly	C. Precision
P. Intercity	32	9	5	0	69.57%
P. Intracity	5	23	1	0	79.31%
P. Highway	5	1	80	21	74.77%
P. Hilly	0	0	11	19	63.33%
C. Recall	76.19%	69.70%	82.47%	47.50%	63.33%

∗C = Class, P = Predicted

It gives overall accuracy of 72.71%. Label is location type for that particular vehicle. Hence algorithm will predict the location type for the vehicles and as per the requirement, suitable vehicle from available dataset will be selected.

V. Conclusion and Future Work

The results are discussed on the comparison of classification of the vehicle using the KNN and NN. The results were observed through the percentage of accuracy of the KNN and NN. In conclusion NN gives better result with backward elimination for feature selection and LOOCV for dataset validation compared to KNN for the system. The proposed system briefly examine the potential use of classification based machine learning techniques to solve industry problem of vehicle selection.

Future work includes the improvement of the system’s performance as well as other modules for the transport management system will be added like minimum distance route plan, ERP etc. Also in future maintenance of vehicle will also be considered as an attribute for classifier building.

References

[1]. M Factors affecting transportation, by NPTEL.

[2]. Machine learning methods for vehicle predictive maintenance using offboard and on-board data, Halmstad University. SerhatAydn , CengizKahraman,

[3]. Vehicle selection for public transportation using an integrated multi criteria decision making approach: A case of Ankara, Journal of Intelligent Fuzzy Systems 26 (2014) 24672481, DOI:10.3233/IFS-13091, IOS Press

[4]. Visalakshi, S., Radha, V. (2014). A literature review of feature selection techniques and applications: Review of feature selection in data mining. 2014 IEEE International Conference on Computational Intelligence and Computing Research. DOI:10.1109/iccic.2014.7238499

[5]. Feature selection guidelines, https://www.machinelearningplus.com/ machine-learning/feature-selection/, [Accessed on :2018-11-10]

[6]. Vehicles selection guidelines, http://www.nzdl.org/gsdlmod?e=d00000-00—off-0fnl2.2–00-0—-0-10-0—0—0direct-10—4——0-1l–11-en-50—20-about—00-0-1-00-0–4—-0-0-11-10-0utfZz-800cl=CL1.4d=HASH31111503a8027f96d6931c.9.3gt=1, [online], [Accessed on :2018-08-12]

[7]. MachineLearning:ValidationTechniques”https://dzone.com/articles/machinelearning-validation-techniques”,[online],[Accessed on :2018-08-12]

[8]. A Quick Introduction to K-Nearest Neighbors Algorithm ”https://medium.com/@adi.bronshtein/a-quick-introduction-to-knearest-neighbors-algorithm-62214cea29c7”, [online], [Accessed on :2019-02-01]

[9]. NEURALNETWORKS”https://www.doc.ic.ac.uk/nd/surprise˙96/journal vol4/cs11/report.html”, [online], [Accessed on :2018-08-12]