Gurukul International Multidisciplinary
Research Journal (GIMRJ)*with*
International Impact Factor 8.249
Peer Reviewed Journal
https://doi.org/10.69758/RXKR8732

e-ISSN No. 2394-8426
Special Issue On Advanced Computational Techniques:
Emerging Trends from Postgraduate Studies
Issue–I(VI), Volume–XII

Prediction of Selling Price of Various Cars

[1]Mr. Vibhav Sathe, [2]Prof. Prerna Dangra
[1]PG Scholar, [2]Assistant Professor
Department of Master of Computer Application,
G H Raisoni University, Amravati, India

Received on: *11 April ,2024* ***Revised on:*** *26 May ,2024,* ***Published on:*** *01 June ,2024*

Abstract: - The manufacturer sets the price of a new car in the industry, with the government incurring some additional expenditures in the form of taxes. Customers purchasing a new car may thus be sure that their investment will be worthwhile. However, due to rising new car prices and buyers' financial inability to purchase them, used car sales are increasing globally. As a result, a used car price prediction system that efficiently assesses the worthiness of the car utilizing a range of factors is required. The current system comprises a system in which a dealer decides on a price at random and the buyer has no knowledge of the car or its current worth. In reality, the seller has no clue what the car is worth or what price he should charge for it. To address this issue, we have devised a highly effective model. Regression algorithms are employed because they produce a continuous value rather than a classified value as an output. As a result, rather than predicting a car's price range, it will be feasible to estimate its real price. A user interface has also been created that takes input from any user and shows the price of a car based on the inputs.

IndexTerms - Used Car Price Prediction, Regression Algorithms, Machine Learning, Linear Regression, Ridge and Lasso Regression, Bayesian Ridge Regression, Decision Tree, Random Forest, XG Boost, Gradient Boosting.

I. INTRODUCTION

Predicting car prices is a crucial task in the automotive industry, involving analyzing factors such as make, model, age, mileage, condition, and market trends. Accurate price predictions benefit buyers and sellers by enabling informed purchasing decisions and setting competitive prices. The importance of predicting car prices has grown with the rise of online marketplaces and data science advancements. Machine learning models can be used to power these predictions, learning from historical data and identifying patterns between features and the final selling price. This study aims to develop a robust model for predicting car prices using various data sources and machine learning techniques. The model will explore variables contributing to car prices and assess their impact, aiming to create a reliable and accurate pricing tool for real-world scenarios. Understanding the dynamics of the used car market and external factors like economic conditions and seasonal trends will also be crucial. The study will detail data collection, preprocessing methods, machine learning models, and evaluation metrics to provide valuable insights and a functional tool for car price prediction.

II. RELATED WORK

Predicting the selling price of cars is a widely researched topic in machine learning and data science. Numerous approaches and methodologies have been explored, including regression models, machine learning models, neural networks, hybrid models, feature engineering and selection, datasets, data sources and preprocessing, evaluation metrics, and recent trends. Linear regression is one of the most widely used methods for predicting car prices, while polynomial regression extends linear regression by considering polynomial relationships between features and the target variable.

Machine learning models include Decision Trees and Random Forests, Gradient Boosting Machines (GBM), Support Vector Machines (SVM), neural networks like Forward Neural Networks (FNNs), deep learning models like convolutional neural networks (CNNs) and recurrent neural networks (RNNs), hybrid models, feature engineering and selection, datasets like the UCI Machine Learning Repository's Car Evaluation dataset, and real-time prediction systems.

Feature engineering and selection involve incorporating features specific to cars, such as engine size, horsepower, brand reputation, and market trends, and reducing dimensionality using techniques like Principal Component Analysis (PCA). Data sources and preprocessing include publicly available datasets, handling missing data, normalization and scaling, and evaluation metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared.

Evaluation metrics include Common Metrics (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared, as well as Cross-Validation techniques like k-fold cross-validation. Recent trends include Explainable AI (XAI) and real-time prediction systems that can predict car prices in real-time using streaming data.

Example studies and applications of machine learning models in predicting car prices include "Predicting the Price of Used Cars using Machine Learning Techniques," "A Comprehensive Study on Used Car Price Prediction Using Machine Learning," and "Used Car Price Prediction System using Machine Learning Techniques." These works provide valuable insights into the methodologies, challenges, and advancements in the field of car price prediction.

III. PROPOSED WORK

The proposed work aims to predict the selling price of various cars by following several key steps. These steps include data collection, data preprocessing, feature engineering, model development, model evaluation, and deployment. Data sources include car dealership websites, online marketplaces, and automotive databases, including information on car attributes, historical selling prices, and market trends. Data types include car specifications, usage metrics, regional price variations, economic indicators, and additional features like GPS and backup cameras.

Data preprocessing involves handling missing values, removing duplicates, correcting inaccuracies, normalizing numerical features, and splitting the dataset into training, validation, and test sets. Feature engineering involves feature selection, feature creation, and dimension reduction using techniques like Principal Component Analysis (PCA). Model development involves experimentation with various machine learning algorithms, hyperparameter tuning, and combining multiple models for improved prediction accuracy and robustness.

Model evaluation involves performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared. Cross-validation is implemented to ensure the model generalizes well to unseen data, and residual analysis is used to identify patterns or biases in the model.

Model deployment involves integration into a user-friendly application or platform, ensuring scalability to handle large volumes of data and concurrent user requests. Continuous monitoring and maintenance are necessary to maintain accuracy over time.

Documentation and reporting are provided, detailing data sources, preprocessing steps, feature engineering techniques, model development process, and evaluation results. A detailed report is prepared highlighting key findings, model performance, and potential areas for future work. By following these steps, the proposed work aims to develop a sophisticated and accurate model for predicting car selling prices, benefiting both consumers and sellers in the automotive market.

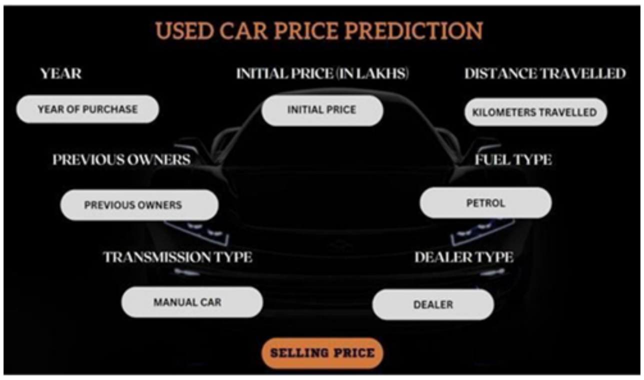

Fig. 1: The flow of proposed work

3.1 Data Collection

Data collection is a critical phase in predicting the selling price of cars, as the quality, quantity, and relevance of collected data directly impact the performance and accuracy of the prediction model. The data collection process involves identifying data sources such as online marketplaces, car dealer websites, automotive databases, and government databases.

The types of data to collect include car specifications, usage metrics, market data, additional features, and data collection methods. Web scraping tools and libraries can be used to extract data from online marketplaces and dealership websites, while APIs provided by services like Kelley Blue Book and Edmunds can access structured data on car prices and specifications. Databases and CSV files can also be obtained and integrated with automotive data. The data collection process includes defining data requirements, setting up data collection tools, collecting data, and storing data in a structured format. Regular checks are made for completeness, accuracy, and relevance. Ethical considerations and compliance include data privacy, data accuracy, and transparency. In summary, data collection is essential for predicting the selling price of various cars, and the quality, quantity, and relevance of collected data directly impact the performance and accuracy of the prediction model. By following this comprehensive data collection plan, a rich dataset can be built for accurate and reliable car price predictions.

3.2 Validation set

A validation set is a crucial step in the machine learning process for predicting car prices. It ensures that the model performs well on both training and unseen data, which is essential for its real-world applicability. The validation set helps in model evaluation, hyperparameter tuning, and prevention of overfitting. The steps involved in using a validation set include data splitting, model training, hyperparameter tuning, model selection, and final evaluation. Data splitting involves splitting the training set into training and validation sets, and evaluating the model on the validation set after each epoch or iteration to monitor performance. Hyperparameter tuning involves using the validation set to tune hyperparameters, such as grid search or random search, to find the optimal combination of hyperparameters. Model selection involves comparing different models and their performance on the validation set to select the best-performing model. Finally, the model is evaluated on the test set to get an unbiased estimate of its real-world performance.

3.3 Testing set

The testing set is a crucial component in machine learning, serving as an independent dataset for evaluating the performance of a predictive model. It is used to assess how well the model generalizes to new, unseen data. The testing set is divided into three parts: the training set, which is used to train the model, the validation set, which is used to tune the model's hyperparameters, and the testing set, which is used solely for evaluating the final model's performance. A good testing set should be representative of the overall data distribution, independent, and sufficient in size. Common metrics used to evaluate the model using the testing set include Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2).

The typical procedure for using the testing set in predicting car prices involves data collection and preprocessing, data splitting, model training, model evaluation, and performance analysis. Data collection and preprocessing involve handling missing values, outliers, and categorical variables, while data splitting divides the dataset into training, validation, and testing sets. Model training involves training the model on the training set and using the validation set for hyperparameter tuning. Model evaluation involves applying the trained model to the testing set to predict car prices and evaluating performance using chosen metrics. Performance analysis determines the model's accuracy and robustness, allowing adjustments if necessary and re-evaluate.

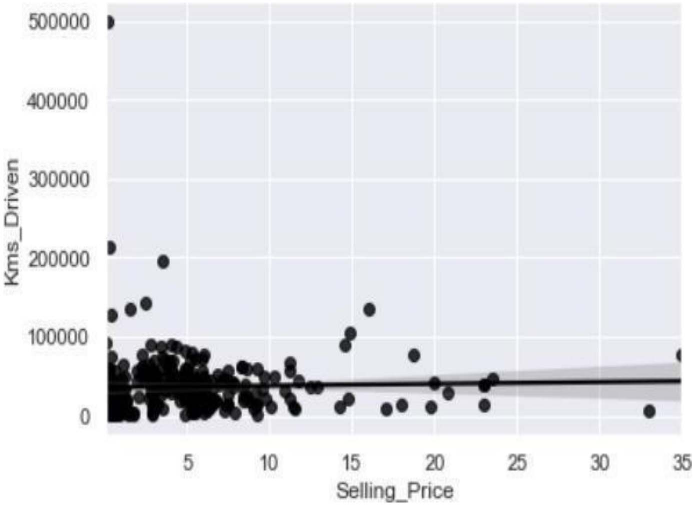

Fig 2.: Initial price v/s Selling price

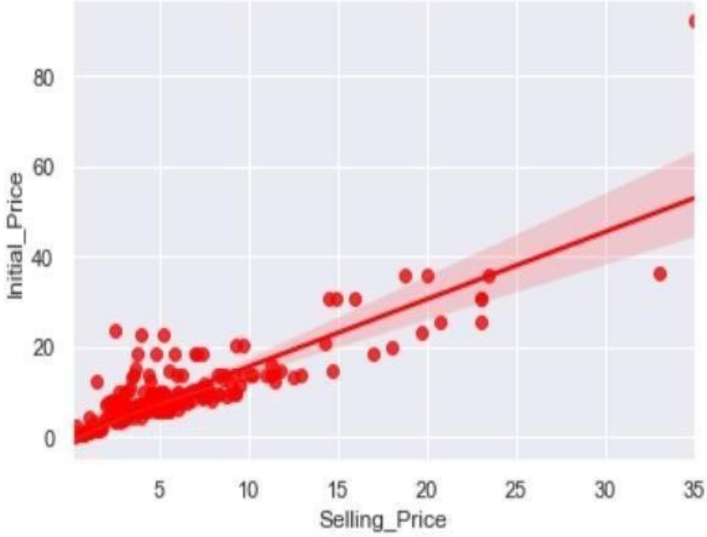

Fig 3. Kilometers Driven v/s Selling Price

IV. RESEARCH METHODOLOGY

The methodology for predicting the selling price of various cars involves a systematic approach that encompasses data collection ,preprocessing, model selection, training, evaluation, and validation. The following steps outline the research methodology:
The process of developing a machine learning model for car price prediction involves several steps. First, a comprehensive dataset is collected from various sources such as online car marketplaces, manufacturers and dealers, public databases, web scraping tools, and APIs provided by car valuation services. Data preprocessing is then performed to remove duplicates, handle missing values, and correct inconsistencies. Feature engineering is also done to improve model performance. Normalization and scaling are also done to ensure all features are on the same scale. Categorical variables are encoded into numerical format using techniques like one-hot encoding or label encoding.

Exploratory Data Analysis (EDA) is used to understand the relationships between different features and the target variable (selling price). Key activities include descriptive statistics, visualization, and correlation analysis. Model selection is made based on the problem and the dataset, with common models being linear regression, decision trees, random forests, Gradient Boosting Machines (GBM), and neural networks.

Model training is done by splitting the dataset into training and testing sets, using techniques like cross-validation and hyperparameter tuning. Performance is evaluated using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2). Model validation is done on a separate validation set to assess performance and prevent overfitting. Finally, the best-performing model is deployed as a web application or API for real-time car price predictions. This involves building a user interface, backend integration, and continuous monitoring and maintenance to maintain accuracy.

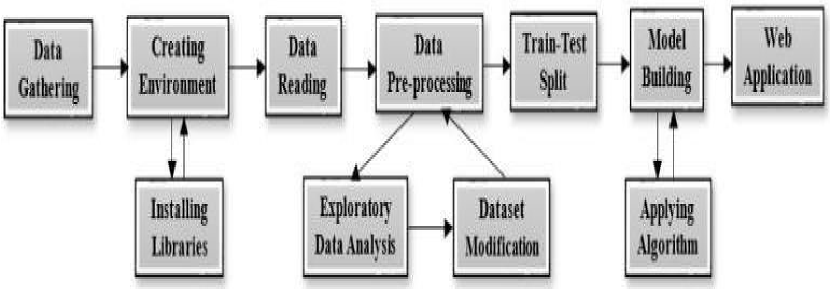

Fig4: Workflow of Study

4.1 Data Pre-processing

Some of the data features were renamed for clarity (Present Price = Initial Price, Owner = Previous Owners), and some features that were not important for analysis were removed. In exploratory data analysis, we use statistical graphics and other visualisation techniques to describe the important aspects of data. Top Selling Vehicles, Year vs. Number of Available Vehicles, Selling Price vs. Initial Price, Vehicle Fuel. Type, Transmission Type, Seller Type, Age, Selling Price v/s Age, Selling Price v/s Seller Type, Selling Price v/s

Transmission, Selling Price v/s Fuel Type, Selling Price v/s Previous Owners, Initial Price vs Selling Price, Selling Price v/s Kilometers Driven, pairplot, heatmaps, and other visualisations are used to gain a better understanding of data. Following EDA, One Hot Encoding approach is used to deal with the dataset's categorical features. After that, the dataset's correlation characteristics are generated and thoroughly analysed by visualising several plots. Then the features allocation of data is where the dependent feature (Selling Price) and independent features (Initial Price, Kilometers Driven, Previous Owners, Age, and so on) are then allocated for further processing.

Train-Test Split: Once the dependent and independent features have been assigned, we proceed with the splitting of the dataset into training and testing data. We use 80% of the data to train our model and 20% to test it.

Model Building Following the Train-Test split, data modeling is complete, and the process of building the model begins. The model is defined, along with a few parameters, for future implementation. After the model is built, various algorithms are used to create the final results. After building the model, the following algorithms are used for predictive analysis

Linear Regression: It is a linear approach in statistics for modeling the relationships between a scalar response and dependent and independent variables. In linear regression, relationships are modelled using functions such as linear predictor, and unknown model parameters are estimated from data.

Lasso Regression: It is a sort of linear regression in which the data values are shrunk towards a data point in the center, or, in simpler terms, the mean of the data. The Lasso procedure supports simple and sparse models with fewer parameters. When a model has a high amount of multicollinearity, this regression provides the best fit for that model. This approach can also be used if some aspects of model selection, such as variable selection or parameter elimination, need to be automated. The abbreviation 'LASSO' stands for Least Absolute Shrinkage and Selection Operator.

Ridge Regression: It is a regression approach used for tuning a model and analyzing multicollinear data. This function implements L2 regularization. The multicollinearity of the data results in unbiased least squares, a huge variance, and hence the predicted values are considerably far from the actual values.

Bayesian Ridge Regression: This regression is used to estimate any probabilistic model of any regression issue using linear regression formulation with the use of probability distributors, providing a natural process that survives data insufficiency or poor data distribution.

Random Forest Regression: Random Forest is a Supervised Learning Algorithm that employs the ensemble learning approach for classification and regression. Random forests are made up of trees that run parallel to each other and have no interaction while they develop. Random Forest is a meta- estimator that aggregates the outcomes of several predictions. It also aggregates numerous decision trees with certain modifications.

XGBoost Regression: XGBoost is a very powerful technique for creating supervised regression models. XGBoost is an ensemble learning strategy that includes training individual models and then merging them (base learners) to get a single prediction.

Gradient Boosting Regression: This is a machine learning approach used to construct a prediction model for regression and classification problems. The prediction model generates an ensemble of weak prediction models, which are often decision trees. This method outperforms the random forest method in most cases.

4.2 Proposed research model

This research aims to develop a predictive model for determining the selling price of various cars using machine learning techniques. The model will analyze multiple factors that influence car prices and generate accurate predictions based on historical data. Key components include data collection, data preprocessing, exploratory data analysis (EDA), model selection, model training and validation, and deployment.

Data collection involves gathering a comprehensive dataset of attributes related to cars, such as make, model, year, mileage, engine size, fuel type, transmission type, color, body type, and features. Market data includes regional market trends, economic indicators, and seasonal effects. Historical prices are previous selling prices of similar cars. Data preprocessing involves handling missing values, removing duplicates, correcting inconsistencies, and converting the raw data into a format suitable for model training.

Exploratory data analysis (EDA) helps understand the relationships between different variables and the target variable (selling price). It involves visualizing distributions of key attributes, identifying correlations between features, detecting outliers, and using statistical measures to summarize the data.

Various machine learning algorithms will be considered for predicting car prices, including linear regression, decision trees and random forests, gradient boosting machines (GBM), support vector machines (SVM), and neural networks. The dataset will be split into training and validation sets to evaluate model performance. Techniques include cross-validation, hyperparameter tuning, and feature selection.

Model evaluation uses metrics such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R-squared (R^2) to compare the performance of different models and select the best one. The final model will be deployed as a web application or API, allowing users to input car attributes and receive price predictions.

```
Data Collection
     |
     v
Data Preprocessing
     |
     v
Exploratory Data Analysis (EDA)
     |
     v
Model Selection
     |
     v
Model Training
     |
     v
Model Evaluation
     |
     v
Model Deployment
     |
     v
Continuous Improvement
```

Fig 5. Proposed Architecture

V. RESULTS AND DISCUSSION

After applying regression algorithms to the model, the r_2 scores and other assessment metrics such as mean absolute error, mean squared error, and root mean squared error were obtained for comparison of the performance of each method. final implementation, the model is created with a few parameters, such as the algorithm, x train, y train, x test, and y test. After the completion of the model, various algorithms are used to generate the final results. The Decision Tree Algorithm has the best r_2 score of 0.9544 when all regression methods' r_2 scores are compared, which simply implies that the Decision Tree Algorithm has delivered the most accurate predictions when compared to the other algorithms.

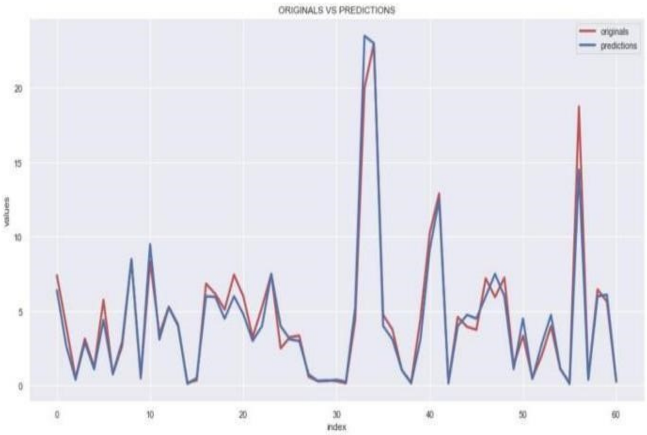

Fig 6. Original v/s Prediction Decision Tree Regression

In the graph above, where the red line represents the original values of the dataset and the blue line shows the values predicted using Decision Tree Regression, we can see that both lines are pretty close to each other, indicating that the predictions are highly accurate.

Algorithm	R_2 Scores	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)
Random Forest Regression	0.8576	0.7583	2.6763	1.6359
Linear Regression	0.8625	1.0998	2.9823	1.7269
Ridge Regression	0.8634	1.1080	2.9632	1.7214
Lasso Regression	0.8659	1.0934	2.9071	1.7050
Bayesian	0.8695	1.0750	2.8302	1.6823

Table 1. Evaluation Metrics of Algorithms

Expected result

The model aims to predict the selling price of cars with high accuracy, closely matching actual market prices using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2). It should identify key features that significantly impact car prices, providing valuable insights into factors determining car value. The model should also demonstrate generalizability, performing well on both training and unseen test data, demonstrating low variance and low bias.

The model should be practical for real-world scenarios, being efficient in computation, easy to use, and integrated into a user-friendly interface. It should account for market variations and trends over time, remaining reliable even when market dynamics shift. The model should be scalable to handle large datasets and a wide variety of car types and attributes, maintaining performance as data increases. The model's performance should be validated through cross-validation techniques to ensure reliability and robustness across different data subsets. By achieving these results, the model will not only provide accurate and reliable car price predictions but also offer valuable insights into factors influencing car prices, aiding buyers, sellers, and dealers in making informed decisions.

VI. CONCLUSION

Predicting used car prices is a difficult task due to the large number of features and parameters that must be examined in order to get reliable findings. The first and most important phase is data collection and preprocessing. The model was then defined and built in order to implement algorithms and generate results. After executing various regression algorithms on the model, it was concluded that the Decision Tree Algorithm was the top performer, with the greatest r2 score of 0.95, implying that it provided the most accurate predictions, as shown by the Original v/s Prediction line graph. Aside from having the highest r2 score, the Decision Tree also had the lowest Mean Square Error (MSE) and Root Mean Square Error (RMSE) scores, indicating that the errors in predictions were the lowest of all and that the results obtained were very accurate.

VII. FUTURE SCOPE

The developed machine learning model can be exported as a "Python class" and deployed as an open source, ready-touse price predictor model, which can then be easily integrated with third-party websites. The model can be greatly optimised by using neural networks by designing deep learning network topologies, employing adaptive learning rates, and training on data clusters rather than the entire dataset.

VIII. REFERENCES

[1] Pudaruth, S. (2014) 'Predicting the Price of Used Cars using Machine Learning Techniques', International Journal of
Information & Computation Technology, 4(7), pp. 753–764. Available at: http://www.irphouse.com.

[2] Kuiper, S. (2008) 'Introduction to Multiple Regression: How Much Is Your Car Worth?', Journal of Statistics Education,
16(3). doi: 10.1080/10691898.2008.11889579.

[3] Pal, N. et al. (2019) 'How Much is my car worth? A methodology for predicting used cars' prices using random forest',
Advances in Intelligent Systems and Computing, 886, pp. 413–422. doi: 10.1007/978-3- 030-03402-3_28.

[4] Gegic, E. et al. (2019) 'Car price prediction using machine learning techniques', TEM Journal, 8(1), pp. 113–118. doi:

10.18421/TEM81-16.

[5] Dholiya, M. et al. (2019) 'Automobile Resale System Using Machine Learning', International Research Journal of Engineering and Technology(IRJET), 6(4), pp. 3122–3125.

[6] Richardson, M. (2009) Determinants of Used Car Resale Value. The Colorado College

[7] Listiani, M. (2009) Support Vector Regression Analysis for Price Prediction in a Car Leasing Application, Technology. Hamburg University of Technology.

[8] https://www.jigsawacademy.com/popular- regressionalgorithms-ml/

[9] https://www.simplilearn.com/10-algorithms- machinelearning-engineers-need-to-know-article

[10] https://www.javatpoint.com/machine-learning-lifecycle

[11] https://www.simplilearn.com/tutorials/machinelearning-tutorial/machine-learning-steps