

TIEVS – DYNAMIC CAR PRICE PREDICTION SYSTEM

2021-195

Project Proposal Report

K.S.S Bandaranayake – IT18113532

Supervisor – Ms. Manori Gamage

Co-Supervisor – Ms. Suriyaa Kumari

B.Sc. (Hons) Degree in Information Technology Specializing in
Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology
Sri Lanka

March 2021

TIEVS – DYNAMIC CAR PRICE PREDICTION SYSTEM

2021-195

Project Proposal Report

K.S.S Bandaranayake – IT18113532

Supervisor – Ms. Manori Gamage

Co-Supervisor – Ms. Suriyaa Kumari

B.Sc. (Hons) Degree in Information Technology Specializing in
Information Technology


Department of Information Technology

Sri Lanka Institute of Information Technology
Sri Lanka

March 2021

DECLARATION

I declare that this is my own work, and this proposal does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any other university or institute of higher learning, and to the best of my knowledge and belief, it does not contain any material previously published or written by another person except where the acknowledgment is made in the text.

Name	Student ID	Signature
K.S.S Bandaranayake	IT18113532	

The above candidate is carrying out research for the undergraduate Dissertation under my supervision.

Signature of the supervisor:

Date:

ABSTRACT

This proposal is prominently focused on implementing a dynamic price prediction system, targeting the items being on sale within our 'Tievs' online classified advertising platform, specifically on Cars. Used Car population has been growing exponentially at an increasing rate due to comparative affordability and leasing culture being widely adopted. Occasionally, Car sellers can be troubled in finding the accurate Car price, leading them to finalize unrealistic costs incurred for demand, making the buyer insecure as well. Hence the significance of a precise price prediction system arises, as it becomes a high research interest area with undertaken expert knowledge and effort. A suitable number of distinct car features that affect the price will be taken into account for more accurate and dependable results. The proposed system, which builds a model for US car price prediction, is based on the combination of three machine learning techniques, namely, LightGBM, Random Forest, and Support Vector Machine. Nevertheless, the models will be trained individually, according to selected features' impact on price, through exploratory data analysis, with respect to the chosen dataset, until an excellent accuracy is received and analyzed. Then, the three will be working as an ensemble model, that of which the precision will be evaluated furthermore, using a test dataset, to produce the final output. The final ensemble model will be integrated with the implemented Angular application to enable the visualization of the predicted price of certain reselling Cars prior to submission of the advertisements the advertisers intend on publishing. As implied, the hypothesis of this novel ensemble model, being more advanced in accuracy, reliability, and inefficiency than existing prediction models, will be exercised while experimenting with the performance of LightGBM for Car price prediction.

Key Words – Car price prediction, LightGBM, Random forest, Support vector machines, Machine learning.

TABLE OF CONTENTS

Declaration	i
Abstract	ii
Table of Contents	iii
List of Figures	v
List of Tables.....	vi
List of Abbreviations.....	vii
1 Introduction	1
1.1 Background	1
1.2 Literature Survey.....	6
1.3 Research Gap	9
1.4 Research Problem.....	12
2 Objectives.....	15
2.1 Main Objective.....	15
2.2 Specific Objectives.....	16
3 Methodology	17
3.1 Research Methodology.....	17
3.2 System Overview	18
3.3 Summarized Steps of System Implementation	20
3.3.1 Data Collection, Storage and Preprocessing	20
3.3.2 Exploratory data analysis	21
3.3.3 Train the selected three models individually	22
3.3.4 Develop the hybrid model and analyze the accuracy level	24
3.3.5 Integration of the model with the application and retraining.....	25
3.4 Software Development Life Cycle.....	26

3.5 Project Requirements	27
3.5.1 Functional requirements.....	27
3.5.2 Non-functional requirements	27
3.6 Gantt Chart	28
3.7 Commercialization	29
3.8 Technologies To Be Used	30
4 Description of Personal and Facilities	31
5 Budget and Budget Justification	32
Reference List	33
Appendix	36
A: Plagiarism Report.....	36

LIST OF FIGURES

	Page
Figure 1.1– Number of Registered Cars from 2003 to 2014	2
Figure 1.2 – Vehicles Operating in US from 1991 to 2019	3
Figure 1.3 – Projection of Chinese Vehicle Growth	3
Figure 1.4 – New and Used Light-Duty Vehicle Sales in US.....	4
Figure 1.5 – Unrealistic Price Value Posted	12
Figure 1.6 – When to retrain the prediction model	14
Figure 3.1 – Research Methodology Diagram	17
Figure 3.2 – System Overview Diagram.....	18
Figure 3.3 – Average price for specific vehicle type	21
Figure 3.4 – Distribution of vehicle age	21
Figure 3.5 – LightGBM Housing Price Prediction	23
Figure 3.6 – Ensemble Model Implementation.....	24
Figure 3.7 – Model Retraining Process.....	25
Figure 3.8 – Agile Methodology	26
Figure 3.9 – Gantt Chart.....	28
Figure 3.10 – Commercialization.....	29

LIST OF TABLES

	Page
Table 1.1 – Comparison of previous researches	11
Table 5.1 – Budget	32

LIST OF ABBREVIATIONS

Abbreviation	Description
SVM	Support Vector Machine
kNN	K-nearest Neighbors
MLP	Multilayer perceptron
MAE	Mean Absolute Error
RF	Random Forest
LightGBM	Light Gradient Boosting Machine
XGBoost	Extreme Gradient Boosting
UI	User Interface

1 INTRODUCTION

1.1 Background

With information technology developments, the usage of online classified advertising portals has been growing immensely throughout the recent years due to many factors such as the higher availability than traditional classified deliverances (newspapers, magazines, booklets, leaflets tend to get inaccessible during a lockdown) [1], ease of use, distributed-architectures, economical, less consummation of precious time out of people's busy schedules and having access to a wider range of customers and sellers. Some of the examples are Quikr, Olx, Craigslist, Carewale, CarDheko, ikman.lk, Facebook marketplace, etc. Nevertheless, some scenarios can occur when customers are not quite capable of finalizing a selling price value for the product there are marketing and presumably will continue on to perform extensive research on how to perceive and clarify a satisfactory price value. When engaged in price clarification, a huge amount of effort must be put into the process, all the while sacrificing valuable time yet again, and in doing so, the customers might feel rather lethargic to continue on with the sale of the product altogether. This withdrawal idea will consequently result in a degradation of satisfying the customer's main requirements and necessities delivered by the particular classified advertising application.

Although our Tievs application will be designed and developed to cater to numerous types of classifieds in the future, along with more upgrades relevant to each criterion, on behalf of the scope of this project that is to be completed within 12 months, we decided on restricting our research implementations, with relevance to only 'Car' classifieds. The population of used or second-hand cars in the world market have been increasing ever since and might have been doubled in the last few years, especially due to the Covid-19 pandemic, and since the traditional ways of trading modes have been unable to meet the demands of the consumer, the online trading platform succession trend has been inevitable for resale of vehicles altogether. Initially, for brand new cars, the price is incurred by its manufacturer and the government, in

addition to various taxing schemes, which then leads to a higher price than most people cannot effortlessly afford due to insignificant amount of funds, resulting in a global increase of used cars, which are comparatively modestly priced. According to data gathered and received from National Transport Authority in 2014 [1] there has been a 254% increase in the number of cars from 2003 to 2014, as shown in Figure 1.1, which is a tremendous variation itself, from 68,524 to 160,701. It infers that brand new cars only acclaim a limited space, out of all cars sold each and every year.

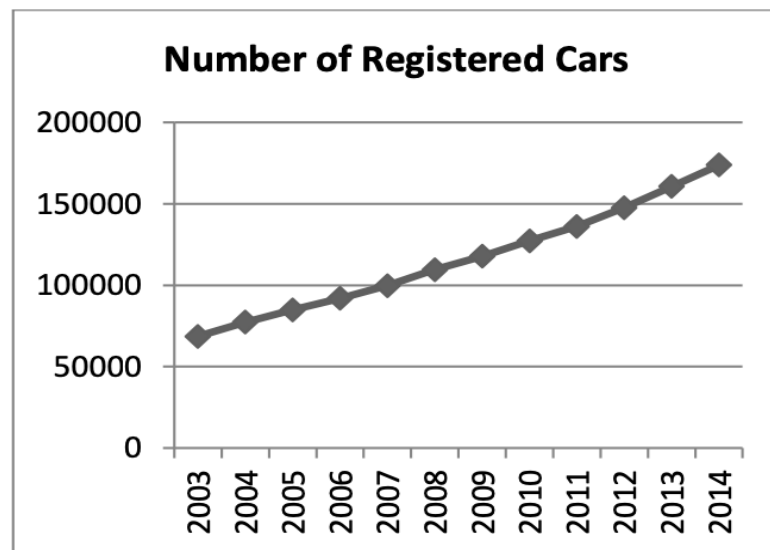


Figure 1.1– Number of Registered Cars from 2003 to 2014

As shown in Figure 1.2, which was acquired from WOLF STREET [2], in addition to Figure 1.3, which was acquired from ‘ACCESS’ online magazine [3], when we carefully analyze the data of the number of passenger vehicles currently in operation in the United States, and the projection of the vehicles in China in the past years and upcoming years, it is plausible to think that the worldwide trend takes a similar form and has continued unceasingly, and in an increasing rate at that as well, regardless of the “Great Recession” period shown in Figure 1.2.

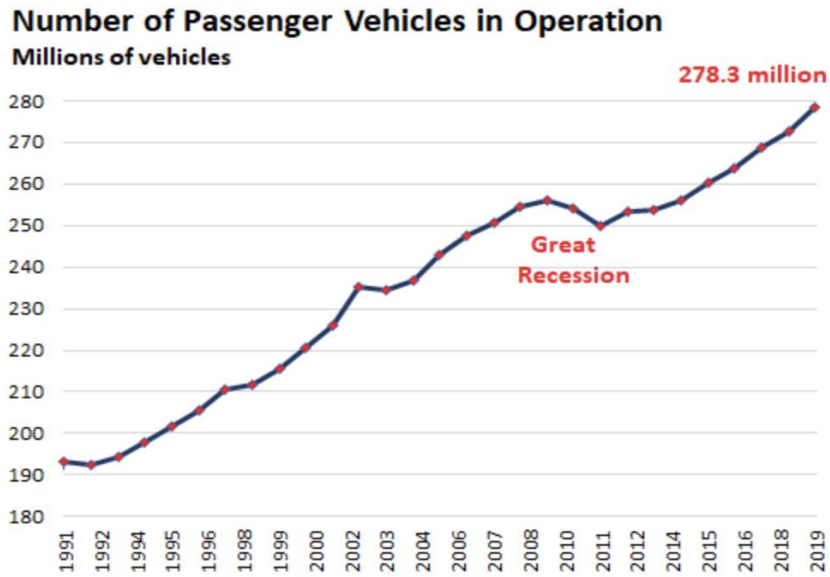


Figure 1.2 – Vehicles Operating in US from 1991 to 2019

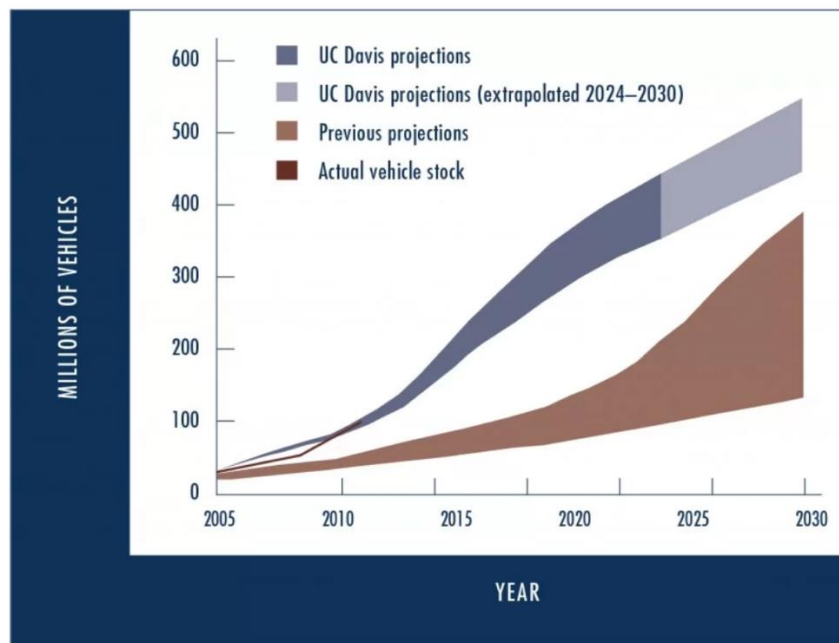


Figure 1.3 – Projection of Chinese Vehicle Growth

Used light-duty vehicle market has progressed to be more than twice the size than new light-duty vehicles within the US, as the statistic variation stands at being 23 million between the two types [4], which figure 1.4 clearly depicts as well.

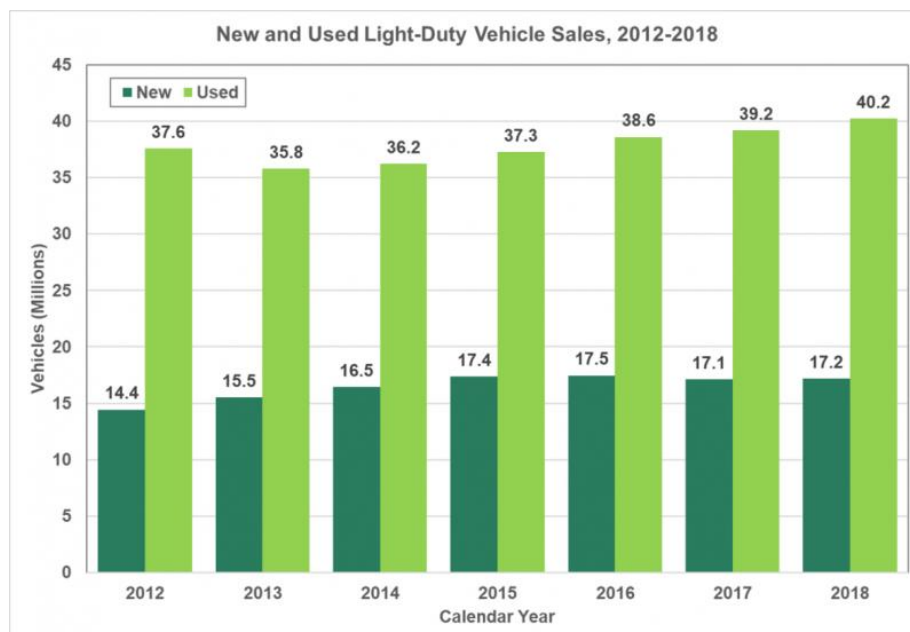


Figure 1.4 – New and Used Light-Duty Vehicle Sales in US

Many cars at present are bought on a lease basis, and after the seller and buyer contract is completed, the buyer is able to resell the car at his or her own discretion. As customers can be lured into impractical prices and fall into misconceptions, it is affirmative that an effective and efficient used car price prediction system is an undeniable necessity to establish accurate worthiness for a car, without any unnecessary inconveniences. Hence, it is why that our Tiefs application should include implementation of an internal process to predict the price of a product, that a certain customer might decide on advertising, and suggest the forecasted price value as a suggestion so that the customer may have the privilege of either accepting and proceeding further with it or simply ignoring it. In this manner, the customer would be more appreciative and eager to stay and utilize the application expedite, without disconnecting from it, due to a single but troublesome matter such as finding out the price value themselves.

As important and seamless as it is, the predicting process of the resale prices of the used or second-hand cars is a tedious and challenging task to undergo, with a high accuracy level, when considering the variety of features and characteristics a car might possess, that promotes deviations in the resale price. Some of the examples are

manufactured year, model, body type, mileage, transmission, braking system, interior type, cylinder volume, fuel max liter capacity, number of doors, size of the car, number of wheels, steering wheel, air conditioner, engine capacity, GPS navigators, overall status, manufacturer awards, car color, braking system, interior designs, acceleration, safety index, car weight, previous owners and types (if the car had gone through any accidents or was it a lady-driven car), and due to the rise of the fuel prices all around the world, the fuel economy and the type of fuel plays a major role in predicting a price for a second-hand car as well, although most people are not aware of the actual amount of fuel that is burned for each km driven by their car. Unfortunately, more or less of these features are considered when people tend to buy a vehicle, and certainly, only a few of them directly and largely affect the price value.

In this component, more characteristics of cars will be considered, and the most impactful ones will be identified. The prediction will be made by using the most optimal machine learning model developed, out of Random Forest Regression, SVM, and Light Gradient Boosted Regression, through thorough finetuning parameters and training the models, which then will ultimately result in an ensemble model type, while using the 'Used Cars Dataset' obtained from Kaggle, which contains prices and attributes of 450,000 of US reselling car details that are updated within the past three months by Austin Reese and extracted from the online classified advertising platform 'Craiglist' by using a web scraper that monitors for data monthly. The dataset will ultimately be preprocessed, and only the relevant columns that make a contribution to the variation for the selling price will be taken into account. The respective achievements and performance of each of the algorithms will be analyzed, with respect to the dataset and the best modal having an inconsequential absolute mean absolute error and other considerable parameters having an exemplary value, will be chosen to be used and integrated within the Tiev's application.

1.2 Literature Survey

Several types of research have been conducted in recent years about this particular car price prediction subject, and there were some phenomenal, revealing results about the accuracy and efficiency levels of machine learning algorithms that have been used for prediction.

Pudaruth [5] has predicted the price for Toyota and Nissan cars in Mauritius by manually collecting the necessary historical data from daily newspapers in less than a month's period since time factors also play an important role in the car price. Attributes like the model, cubic capacity, mileage, production year, exterior color, brand, transmission type, and price of a car were studied by him. Four different techniques like multiple linear regression, decision trees (Random Forest and J48, which is a Java implementation of C4.5), k-nearest neighbors, and Naïve Bayes were compared by the author to finding out the best model. The data were normalized to prevent large values when used with kNN, and numerical values were converted to nominal values to be used with decision tree algorithms. It was concluded that Naïve Bayes and decision tree methods were unable to handle numerical output classes while kNN obtained the best results having a mean absolute error of 27,000 for Nissan cars and 45,000 for Toyota cars. Due to the limited number of data records (97) considered, the performance accuracy was not at an excellent level, i.e., being less than 70%.

Noor and Jan [6] used a supervised machine learning technique, a multiple linear regression model, to predict the price of a car, and their dataset was collected during a two-month period, having car attributes such as model, cubic capacity, mileage, manufactured year, transmission type, number of ad views, brand, exterior type, power steering, rims type, engine type, city, registered city, version, date when the ad was posted, and price. They collected 2000 car data records having the prices of them tagged, from a famous Pakistani car reselling online platform called PakWheels, and preprocessed the data, which was then finalized to 1699 records. The authors applied variable selection techniques to find only the most relevant features from the list, such as model, model year, price, and engine type. Since regression is based on numerical

data, text values in columns as engine type and model were converted to codes, i.e., 1,0, and then fed into the model using Minitab. In conclusion, the authors were able to achieve an impressive R-sq value of 98.61% and R-sq (adj) value of 98.50%, since depending only on R-sq is impractical due to natural variability when adding more predictors into the model and contaminate the pure accuracy by making R-sq value higher without solid verifications.

Similar to the research done by Pudaruth [5], Peerun et al. [7] has also conducted research on predicting the price of Mauritian reselling cars by comparing the performance of four machine learning algorithms such as SVM, kNN, Linear Regression and MLP (cycles = 5000, learning rate = 0.5). They have gathered 200 record dataset through car websites and daily newspapers, during an interval of one month (month of August 2014), having car features as the year, make, paint, engine, mileage, and transmission type, along with the specific prices given. The authors' experiment results depict that SVM and neural network regression (MLP) had slightly better performance than linear regression when comparing the accumulated MAE (in Rupees) values which are, respectively, 30605 for SVM, 30746 MLP, 30828 for linear regression. kNN had the largest MAE, which was 42240, and considered the worst accurate of the four approaches. Even though the MAE values are higher for these algorithms than the values obtained by Pudaruth [5], the authors have considered a higher amount of records (200) as opposed to 97 records, and have considered six predictor factor than just three factors, making the prediction somewhat satisfactory since the deviation is only less than 10% from the actual price.

Another research was done to compare the algorithms Random Forest, ANN, and SVM and finding out the most reliable and accurate predict model, with regard to the 797 records preprocessed car dataset, initially obtained from autopijaca.ba web portal during the winter season in Bosnia and Herzegovina, using a PHP web scraper. Authors Gegic et al. [8], however, made an ensemble modal out of the three algorithms and proved the fact that rather than opting for a single algorithm from the above stated three, which each of them individually had less than 50% accuracy, a hybrid version employs much better potential in predicting the price with 92% accuracy.

According to the authors' Sun et al. [9], the optimized algorithm they have built, called the Like Block-Monte Carlo Method (LB-MCM), by basing BP neural network algorithm using 1630 records of data, identifies hidden neurons more swiftly, having both generalization and approximation ability. The authors infer that the built model will increase convergence speed, accuracy, and network topology than traditional formula methods while providing price prediction for a P2P car trading system. The absolute and relative error rates encountered prior to tweaking the BP neural network algorithm was respectively 0.113 ± 0.080 and $0.78\% \pm 0.55\%$, while after the optimization was done, the values were reduced to 0.084 ± 0.069 and $0.58\% \pm 0.48\%$, hence the immense robustness, strong accuracy, and higher fault-tolerance, as claimed by the authors themselves.

Author Listiani [10] has used SVM to predict leased car prices and showed a better accuracy rate than multivariate or a simple multiple linear regression when a large dataset with more dimensions is present, which also reduces overfitting and underfitting. The downside of it was the absence of indicators as mean, standard deviation, and variance to show the changes between multiple linear regression and SVM regression. Through the research done by authors Pal et al. [11], they have shown the high performance in predicting car prices using Random Forest regression, using 500 decision trees, and exploratory data analysis to find the impact of car features on price. They have used the Kaggle dataset 'Used Car Database' with 370,000 records and 20 car attributes, although after preprocessing, they considered kilometer, vehicle Type, brand, and price as most prominent. The authors received 95.82% for training accuracy and 83.63% for testing accuracy from RF.

Likewise, there are fewer researches prevailing that have considered a dataset with a high amount of records, and none of them have considered LightGBM regression for predicting a car price, while most considered individual algorithms for prediction. Hence, I have decided to take into account LightGBM, and two other machine learning algorithms identified through the Literature review, having the excellent capability, make the necessary refinements, analyze individual model accuracy and implement a hybrid model to predict the price, considering composite fidelity and time complexity.

1.3 Research Gap

Even though it is perspicuous that some researchers have attempted to implement an outstanding model that is capable of predicting prices of used or secondhand cars with decent accuracy levels, still much more undiscovered innovative methodologies are yet to be realized.

Most of the research that were conducted before has used only a small number of records as their datasets. For example, Pudaruth [5] considered only 97 records to train and analyze the models, while Noor and Jan [6], Peerun et al. [7], Enis Gegic et al. [8] used 1699, 200, and 797 respectively, which is extremely below than the amount that this proposed component would be working on, which is about 450,000 records (after preprocessing can be reduced but slightly).

Hence the issue of models being biased to training data will be reduced since the range of records is higher and will additionally improve the prediction accuracy when there are more samples as well. On the other hand, some research have considered only a few range of car characteristics for the prediction [5]-[7].

Through this component, further analysis using specific techniques like feature engineering will be done to identify the most relevant features that make an influence to the price, even at the slightest level, and consider those features to be as inputs to the models.

When considering the available research papers [5]-[11], it was vividly noticeable the authors have never used the LightGBM regression model for the price prediction of reconditioned, used, and secondhand cars. Nevertheless, Truong et al. [12] have pointed out that the LightGBM regression model is a successful predictive model itself, having the speed faster than other similar performing solutions in the field, flattering scalability, and low memory usage, when compared with XGBoost, are its invaluable qualities [13].

In the same research by the authors Truong et al. [12], they have concluded that, when predicting the prices of houses, the LightGBM model had the RMSLE value of 0.16687, which is pretty decent in contrast to the other investigated models, but the pinnacle of it was its speed of training. Hence, the proposed component will be utilizing the LightGBM model as one of the models that are being considered.

When analyzing the research papers further [5]-[7], [9]-[11], it is understandable to the reader that the authors have analyzed and investigated the performance, accuracy, and mean error levels, for only a single machine learning algorithm, in the context of car price prediction. RF, ANN, SVM, multiple linear regression, and kNN were the most popular singular executing algorithms chosen for the task by many. However, Gegic et al. [8] have proved that rather than applying a single model, a composition of two or more algorithms works better in predicting the car prices more accurately with better precision. The authors of that research have manipulated three algorithms individually and came to the settlement that the error rates were high, and accuracy levels were less than 50%, although when used as an ensemble model, the accuracy was risen to above 92% and addressed the overfitting problems of certain individual models and improves generalization. Therefore, the proposed system will be implementing a hybrid model with three predictive algorithms that have never been attempted as a combination in previously fulfilled explorations.

Many types of research [5]-[8], [10]-[11] have not given attention to the retraining or re-learning aspect of their model or models under experiment. The discussions were more biased to the initial training of the algorithm models through thoroughly preprocessed datasets, analyzing the final price prediction results through testing, and finally deploying the model, integrated with a frontend application. Hence, the re-learning functionality will be implemented within the proposed ensemble model, as another novelty aspect, to keep abreast of the newest additions into the application by customers regarding car details and characteristics. With time, market prices are continuously changing for cars, hence the need and significance of continuous re-learning of the prediction system itself, with the help of new data, in order to be up to date with the current approximate market values, under the assumption that the

information regarding cars that are being submitted by customers into the Tiev's application, like advertisements, contain genuine and faultless data about cars. Through the retraining approach, the model will perform much better in prediction at large, overcoming the challenge of time. Table 1.1 below depicts in tabular form, the summarization of the above explanation.

Table 1.1 – Comparison of previous researches

Research	Complex dataset having a higher number of records	Consideration of more features affecting the price of a car	Usage of LightGBM model in price prediction	Implementation of a ensemble model having a composition of models to predict the price	Re-training final developed prediction model for continuous accuracy
Research A [4]	×	×	×	×	×
Research B [5]	×	×	×	×	×
Research C [6]	×	×	×	×	×
Research D [7]	×	✓	×	✓	×
Research E [8]	×	×	×	×	✓
Research F [9]	✓	✓	×	×	×
Research G [10]	✓	×	×	×	×
Proposed Prediction System	✓	✓	✓	✓	✓

1.4 Research Problem

It is apparent that the second-hand or used car population has grown higher than the brand new car population having the causes as people's fund management issues and popularity of leasing approaches. Thus, it is a delicate task to decide on an affirmative price that is fair both to the car reseller and the buyer. It is further tedious when the market prize hunting will need to be undergone manually, by the reseller themselves, through extensive researches, while wasting precious time out of their schedules. The below Figure 1.5 shows a real-world example, acquired from Craigslist Classifieds, of a situation where a customer has carelessly posted an unrealistic price (as \$0) for the reselling vehicle.

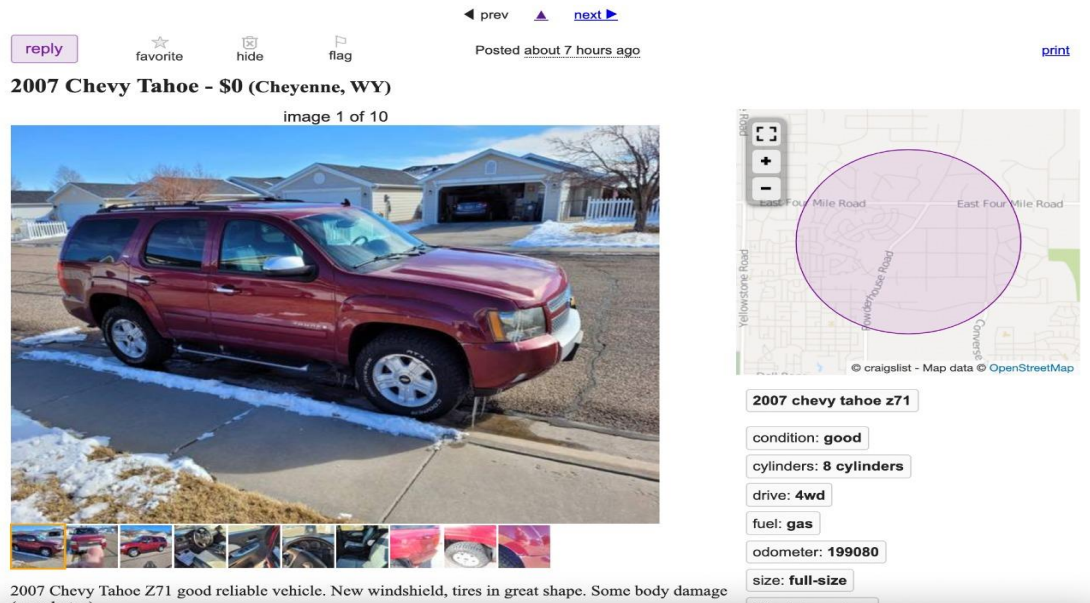


Figure 1.5 – Unrealistic Price Value Posted

To address the issue, experimenters have attempted researches as [5]-[11], to build machine learning models that have the correct intelligence and performance for price prediction of used cars.

However, When analyzing through the researches, many machine learning models designed either from regression or classification procedures have been sophisticatedly investigated, and the results produced have been successful in some scenarios,

although others were not commendably appreciative. Most of them have not considered complex datasets with a large set of records for training and testing of the models, and even though, unquestionably, more data will be making the model higher inaccuracy. Analyzation of features that contributes to the price needs to be extended from four or six features, all the while maintaining simplicity of the model, which was not thoroughly investigated by most researchers, except for [9], in fear of prediction model complexity.

Similar to accuracy, speed is a prominent factor in prediction as well, although the LightGBM regression model, with an uppermost range in speed according to [12], hasn't been explored regarding car price prediction in previous experiments, which is a perceptible gap in research. Proven by authors of [8], using a single model for prediction realizes less accuracy and more error gaps than using a composition of predictive models, which gives superior prediction power in comparison. Regardless, a majority of the experimentations have appraised models individually, by inspecting the desired accuracy levels are being depicted or not, according to dataset deviations and model parameter tweaking, for example, researches such as [8]-[10].

Some have taken into account different machine learning algorithms, for example [4]-[6], when conducting their research studies, and never did opt to apply for the hybrid model and analyze the results to see if the performance increases or not. They compared the performances of the models separately by training them with the same dataset and finally concluded the model having the best accuracy and least error rates as the best price prediction model to be used in the market.

Thus, utilizing ensemble models is obligatory if precision is valuable, though the usage is extremely rare in the field of car price prediction, and the combination of the appropriate models to be considered are plenty for exploration.



Figure 1.6 – When to retrain the prediction model

The above depicted Figure 1.5, which was demonstrated in [14], defines instances, a machine learning model should be trained, and [15] presents the importance of retraining that arises due to factors such as changes over time in the monitoring and learning contexts with the addition of new data and further states that if retraining is neglected, model recognition accuracy will decrease with time. [14] implies that, unlike software applications, machine learning models need continuous training to consistently predict with accuracy, though many types of research like [5, 6, 7, 8, 10, 11] have not argued the significance adequately, hence a major misplacement for persistent price prediction in related works. On that account, the proposed price prediction system intends to address the prevailing problems in equivalent research studies, which are stated above, and sufficiently make novel amendments accordingly as solutions to enhancing the user experience for the customers of the application as a whole.

2 OBJECTIVES

2.1 Main Objective

- The ulterior objective of this proposal is to implement a dynamic price predicting system, within the scope of one year, using a hybrid predictive machine learning model, to recommend a selling price value for specific products (Cars) being advertised on the ‘Tievs’ classified advertising platform.

When diving into the bigger picture, upon the accomplishment of this objective, it will serve with high importance as one of the major fresh inclusions for the overall ‘Tievs’ SMART online classified advertising system, making it surpass the existing applications that publish classifieds. Most of the available systems do not include an internal machine learning model-driven process to predict a price of a product and suggest the customers with it through the application’s user interface. Hence it has become the main objective of this proposal to facilitate ‘Tievs’ with the enhancement so that it astoundingly emerges from other competitors. However, due to scope and time constraints inflicted, the system will implement these functionalities, targeting one type of classified, which is Cars. Subsequently, this component will be focused on predicting the reselling price of the used or second-hand car and displaying it as a suggestion, according to the specifications depicted by the customers on the advertisement form they fill prior to submission. Undoubtedly, the overall population of cars throughout the world has increased when compared to earlier years and still continues to do so [1]-[4]. Hence the reselling rate has become towering due to fewer funds people have at their disposal and opting for a leasing contract between the buyer and seller, where at the completion of the whole value payment, the buyer is capable of reselling the vehicle themselves. In these situations, people who intend to sell their vehicles and the ones who seek to buy could be swindled by fraudsters and confront disregard for their money due to unreasonable prices tagged. Ostensibly, the significance of an accurate, reliable, and efficient price prediction system for reselling cars is portrayed, thus the main objective of this proposal.

2.2 Specific Objectives

In contemplation of achieving the main objective, there are certain sub objectives that need to be addressed fundamentally.

1. Implementation of an ensemble model with a novel combination of algorithm models. In [8], it was implied that the usage of a hybrid model, with a collaboration of two or more algorithms models, can predict the price of a car more accurately than a single machine learning model. However, many ignored attention to the proposition and have only used single models for price predictions [5, 6, 7, 9, 10, 11]. Hence, it has become an objective of this new proposed prediction system to investigate on the precision of an ensemble model, with a new combination of LightGBM, RF, and SVM, and argue about the usability and reliability.
2. Successful utilization of LightGBM algorithm, as a part of the ensemble model, for car price prediction. The proposed system will be evaluating the functionality and sustainability of the LightGBM algorithm model as to validate the investigations and hypothesis of the models speed and scalability, presented by [12]-[13], prompting for a new contribution, in car price prediction research field.
3. Developing the retraining ability of the ensemble model to maintain persistence throughout the life cycle of the application. Along with the various changes incurred within the learning contexts of the model, especially with time, the previously developed model's prediction accuracy may start to decline gradually and would not be completely accurate, hence the need for retraining the model, along with new data acquisitions in the application.
4. Conclusive adaptation of the model, with respect to a complex dataset, having a large number of records and car characteristics for rumination of price prediction, aiming for better accuracy and precision, using more information for model training and testing, resulting in a 'best' model performance.

3 METHODOLOGY

3.1 Research Methodology

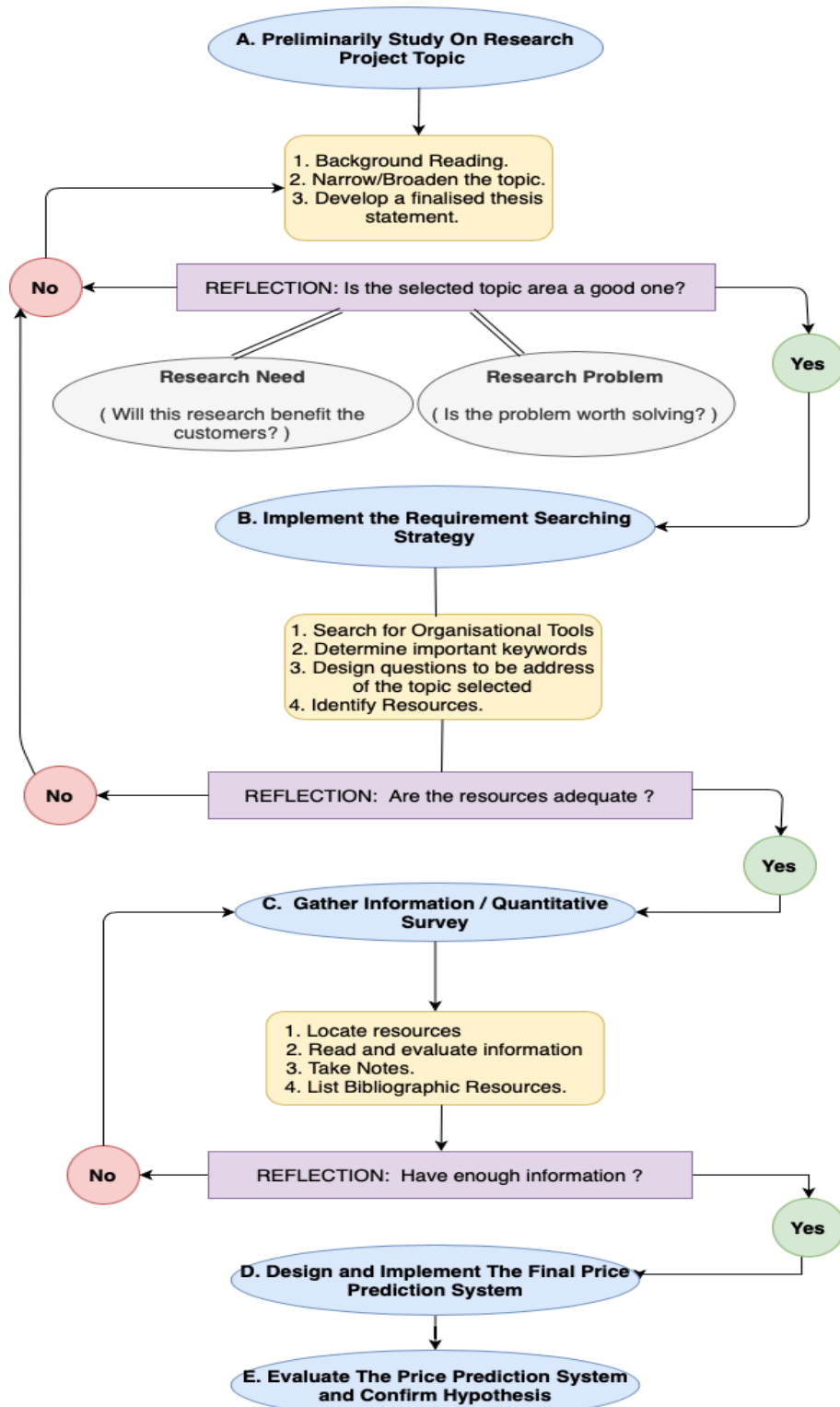


Figure 3.1 – Research Methodology Diagram

3.2 System Overview

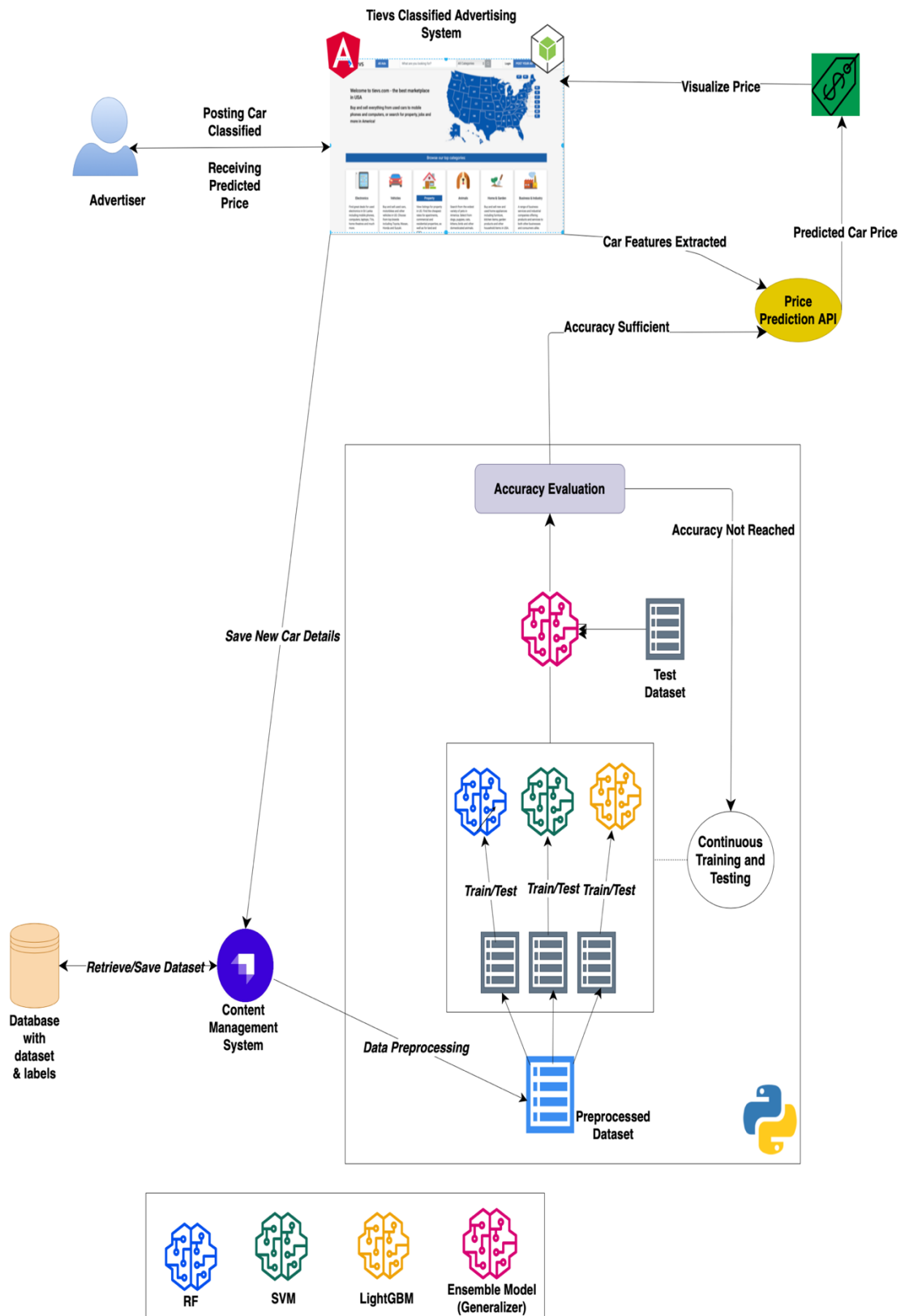


Figure 3.2 – System Overview Diagram

The above-shown Figure 3.2 depicts the overall high-level diagram of the proposed component system to be built for the prediction of car prices. First, the dataset will be requested by the Python machine learning API, through a content management system called ‘Strapi’, from the database where it resides. After acquiring the dataset, various attribute selection techniques, feature engineering techniques, and various data preprocessing methods will be implied to cleanse the dataset prior to transferring it into the models. After the dataset is accordingly prepared, it will be split into subsets and fed on to the three types of machine learning algorithm model types that are being considered for this particular research concept, namely, RF, SVM, and LightGBM.

The model parameters will be adjusted accordingly to the dataset characteristics. The training process will need to be applied repetitively for each of the models for better learning perception. The accuracy level will be analyzed for individual models, and afterward, by using an ensemble method for the model combination, a hybrid classification or a regression model will be implemented. Using the test dataset, the accuracy of the hybrid model will be analyzed, and if refinements are necessary, the process will navigate back to the training of the individual models once more. Subsequently, succeeding the precision of the hybrid model, it will be integrated with the Angular application in order to predict prices of cars for the advertisers who are interacting with the UI. Car features will be extracted from the advertisement form in the application interface into the machine learning API, having the generalizer model, and after meticulous analysis of the features, the appropriate price will be produced. The predicted price will be visualized to the customer, within the advertisement form, near the price input section. New additions being submitted into the application, having car advertisement details, will be saved in the database and will be utilized for the purpose of retraining the prediction model.

For the optimum functionality of the above-described system, the next stated aspects, which are rigorously emphasized, should be thoroughly realized and attained.

3.3 Summarized Steps of System Implementation

3.3.1 Data Collection, Storage and Preprocessing

Since manual data collection on cars would be tedious and time-consuming, this proposed system will be using an available online dataset. The dataset was initially published and updated by Austin Reese, which is scraped from one of the popularly listed online classified advertisement platforms, Craigslist, which is sparsely known and being used by many customers within the United States. The ‘Used Cars Dataset (vehicles.csv)’ conforms to about 450,000 records about sold cars of various brands through the Craigslist website and includes 26 attributes of defining characteristics of cars, that of which some will be excluded due to irrelevancy or empty fields. The reason that this particular dataset is ideal for the implementation of the car price prediction component of the Tiefs application is that the application is fundamentally being built by targeting the consumers of the US. Hence, the information about sold cars in the US itself is appropriate for the progress of this matter in hand. Storage of the dataset will be done by using the MongoDB or the MySQL database environment, which are widely known platforms within the industry for their complimentary performance in web applications. After the dataset is being properly stored, the next step, which is data preprocessing, should occur. It is known as the most challenging and major part of the whole experimentation process, according to [7]. They further mention that the reason for the increase in performance of a prediction model would be the sufficient amount of data preprocessing done exceptionally. For example, some of the preprocessing tasks that can be done on the currently selected dataset would be,

- Removal of columns values having ‘N/A’ (Not Available) as the majority.
- Conversions, categorizations and normalization on certain columns.
- Filtering out the irrelevant columns that does not affect the price of a car. (Ex: Advertisement name, advertisement publisher, pictures, unnecessary URLs, postal codes)
- Excluding the rows having ample amount of empty values.
- Excluding all cars that do not have the associated price value.
- Removal of the records having unrealistic engine power values. Etc.

3.3.2 Exploratory data analysis

In order to get some insight on how the dataset can be utilized along with the models to be analyzed, visually descriptive methods can be implied. It will enable the author of this proposal to better identify the correlations between the existing car characteristics with the relevance to their prices. Basically, a viewpoint can be established on what ranges the attributes in the dataset reside, the diversity of each and every feature in the dataset, and to determine the correlation between each feature considered, including the target feature price itself, significantly vary or stay dense. For the purpose of visualization, bar charts, box plots, scatter plots, distribution tables, etc., could be manipulated, similar to the below-shown diagrams, which have been obtained from [10].

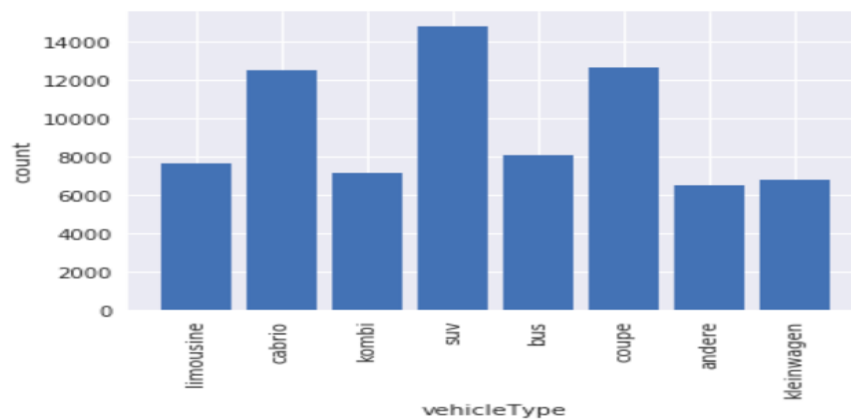


Figure 3.3 – Average price for specific vehicle type

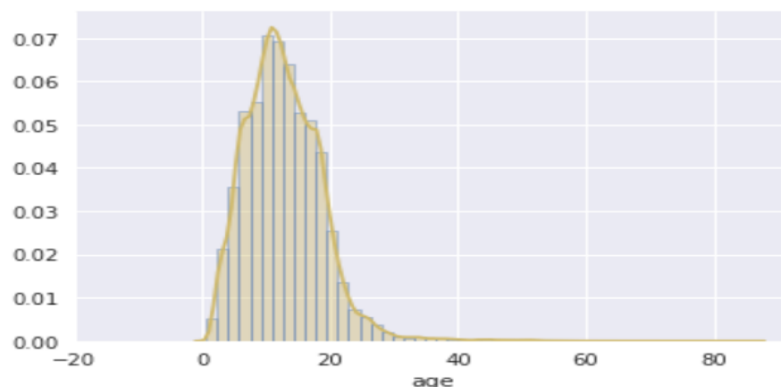


Figure 3.4 – Distribution of vehicle age

3.3.3 Train the selected three models individually

Split the dataset into train and test sets, redefine parameters accordingly to maintain accuracy and speed of the models, which are being considered and analyze the accuracy levels of each of them individually. When tweaking the model parameters, one must be aware of each algorithm's types' capacity of performing, when a certain number of features are taken into consideration and when a complex dataset is imminent. When training, the dataset can first go through a certain algorithm and then after another. In this stage, the dataset can be divided categorically, and each subset type can be applied to each of the models in order to train. Can change the dataset splitting percentage and experiment the results, in addition to monitoring the speed and efficiency of each algorithm model. The chosen machine learning algorithm types are RF, SVM, and LightGBM.

- Random Forest

Random Forest itself can be categorized as an ensemble model. It can be used for the classification or regression type problems and contains a cluster of decision trees, either in hundreds or thousands in number, and goes very deep in training each tree separately [7, 10]. Ho has developed this particular algorithm to address the overfitting problem of decision tree algorithms overall [16].

- Support Vector Regression

This algorithm can be incorporated for solving both classification and regression type algorithms as well. It can be easily trained to tag input data with labels, to divide them in the widest area possible, between categories [17, 16]. In the event of label unavailability, SVM cannot be applied for it. It requires to apply another version of SVM, known as Support Vector Clustering, which is an unsupervised method, to categorize unlabeled data [17, 18].

- Light Gradient Boosting Machine

LightGBM is an inclination boosting structure that utilizes a tree-based learning calculation. LightGBM has a quicker training speed with lower memory use in contrast with XGBoost [12]. Besides, it can likewise uphold equal and GPU learning or handle the enormous size of information and develops tree leaf-wise as well. There are itemized differences in displaying, making them exceptional among various gradient boosting models. For insightful demonstration purposes, the house price prediction accuracy graph of the LightGBM model depicts below, which was initially acquired from the research [12], although the model hasn't been used for car price prediction as of yet.

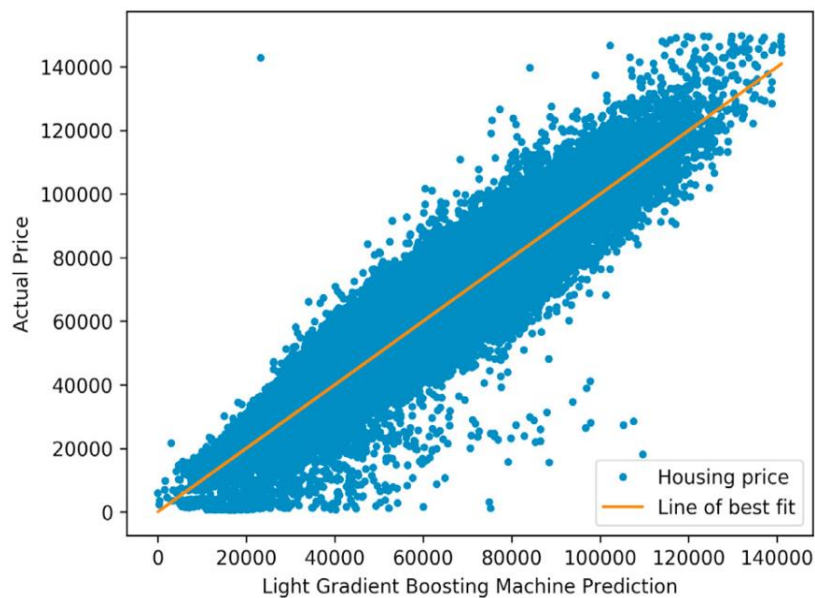


Figure 3.5 – LightGBM Housing Price Prediction

The basic objective is to implement a hybrid model, by combining the above mentioned models, but first, individual analysis among the models will be executed foremostly, as to get a better perception on how the ensemble model's accuracy will differ in comparison to the singular models themselves.

3.3.4 Develop the hybrid model and analyze the accuracy level

Use a suitable ensemble methodology and combine the three algorithms together, that were previously used as single algorithms. The following mentioned techniques can be used to develop models with high accuracy and value [20].

- Majority voting
- Weighted voting
- Simple averaging
- Weighted averaging
- Stacked Generalization
- Bagging
- Boosting

After extensive training of the individual models, they can be assembled into one hybrid model as the final prediction system. Then testing of the hybrid model will be done to clearly and accurately evaluate the authenticity of predicting a price. The below Figure 3.6 shows the abstract view of how the ensemble model will be implemented.

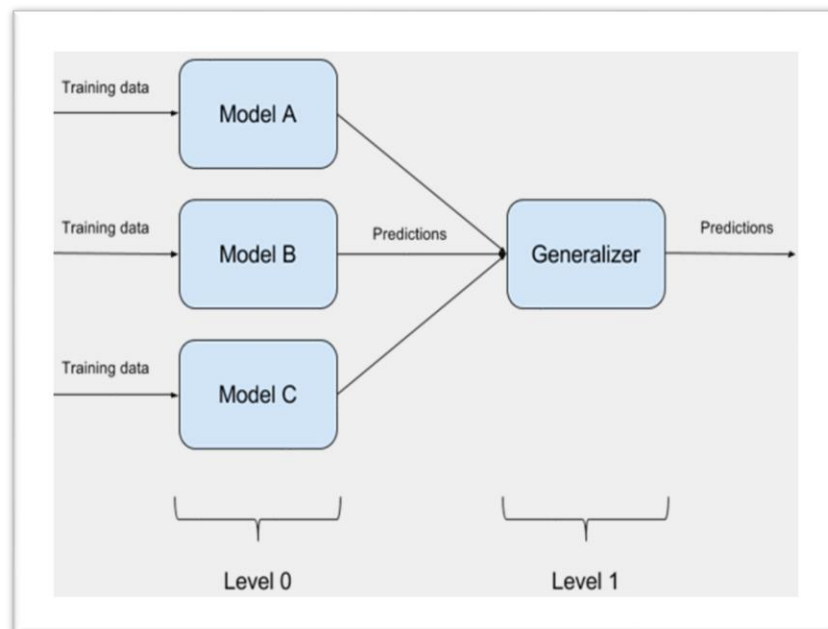


Figure 3.6 – Ensemble Model Implementation

3.3.5 Integration of the model with the application and retraining

The fully trained and tested hybrid price prediction system needs to be deployed on the server where the application is being held and integrated with the necessary front and backend components built from the Angular framework powered ‘Tievs’ application, particularly within a machine learning API, along with the other solutions presented by the rest of the research members, without any occurring errors or inconveniences. With fruitful attainment of this objective, the road to price prediction and recommendation of it to the end-user will be achieved successfully. As the final task to be executed, the implementation of the retraining process of the prediction model, using the batch retraining technique with the help of Jenkins or Kubernetes Cronjobs, will be done as well, so that the prediction accuracy can be maintained further continuously, even though unseen data has been encountered. The following Figure 3.7 depicts the basic processes essential for the retraining aspect of the prediction model to be a reality.

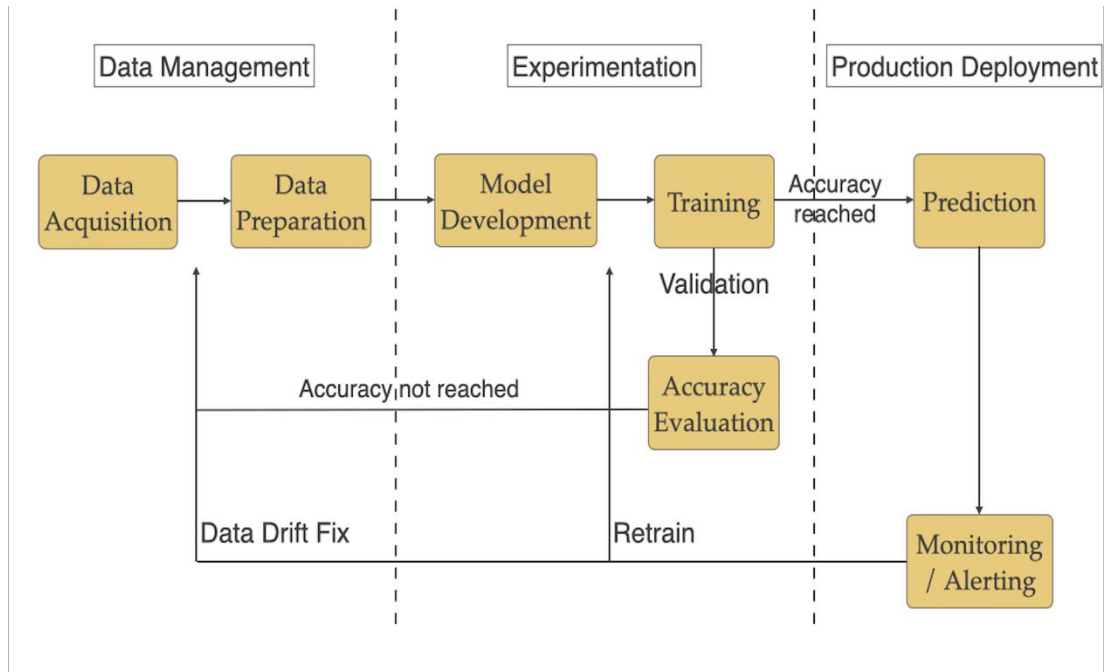


Figure 3.7 – Model Retraining Process

3.4 Software Development Life Cycle

This project will be guided throughout the year, with the infamous continuous delivery and continuous development aspired, change welcoming, methodology approach, Agile, as presented in Figure 3.5, from [21]. Continuous delivery, face-to-face interactions, motivated personnel, change embracement, working software deliverables are some of the main points of consideration in the Agile procedures and those of which would be heeded within the development of this proposed price prediction system. Although a set plan is constructed to aid for the time management and delivery schedule alignments with the help of a Gantt chart, according to the Agile values, if impediments were to occur, the necessary adaptations must be taken into place in a timely manner, without interrupting the continuous development processes [21]. Especially when conducting research, grey areas, with many uncertainties either in the methodology procedures or in the techniques in use, will be encountered, that of which a clear understanding is not being procured, as the literature survey has suggested. At those times, responding effectively and positively to change and maintaining the consistency of the schedule plan is of high significance. If Agile methodology is followed within the software development life cycle, it would be much easier to contemplate the obstacles by making them the opportunities for success itself. In conclusion, Scrum, which is a lightweight version of the vast Agile framework, will be utilized to cater the incremental deliverables with high quality during short periods along with overall team and project management.

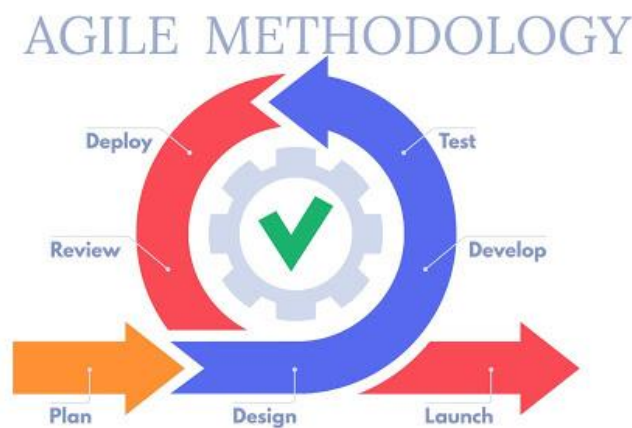


Figure 3.8 – Agile Methodology

3.5 Project Requirements

3.5.1 Functional requirements

- Customers should be able to navigate to the advertisement posting form successfully.
- Correct visualization of the advertisement form components for a convenient depiction of the various characteristics of a car should be implemented.
- Accurate functionality of the ‘Generate Predicted Price’ button should be maintained.
- Extraction of the car characteristics into the prediction model should happen accordingly.
- Precise price prediction according to extracted details supplemented into the prediction model shall be implemented.
- Shall clearly present the predicted price to the customer.
- Should acquire details of newly submitted car classified information.
- Retraining of the ensemble model should be performed.

3.5.2 Non-functional requirements

- Accuracy
- Less response time
- Scalability
- Efficiency
- Effectiveness
- Ease of use
- Time-saving
- Adaptability
- Accessibility
- Maintainability

3.6 Gantt Chart

The Gantt chart of the proposed system's development procedures are shown in the below Figure 3.7.

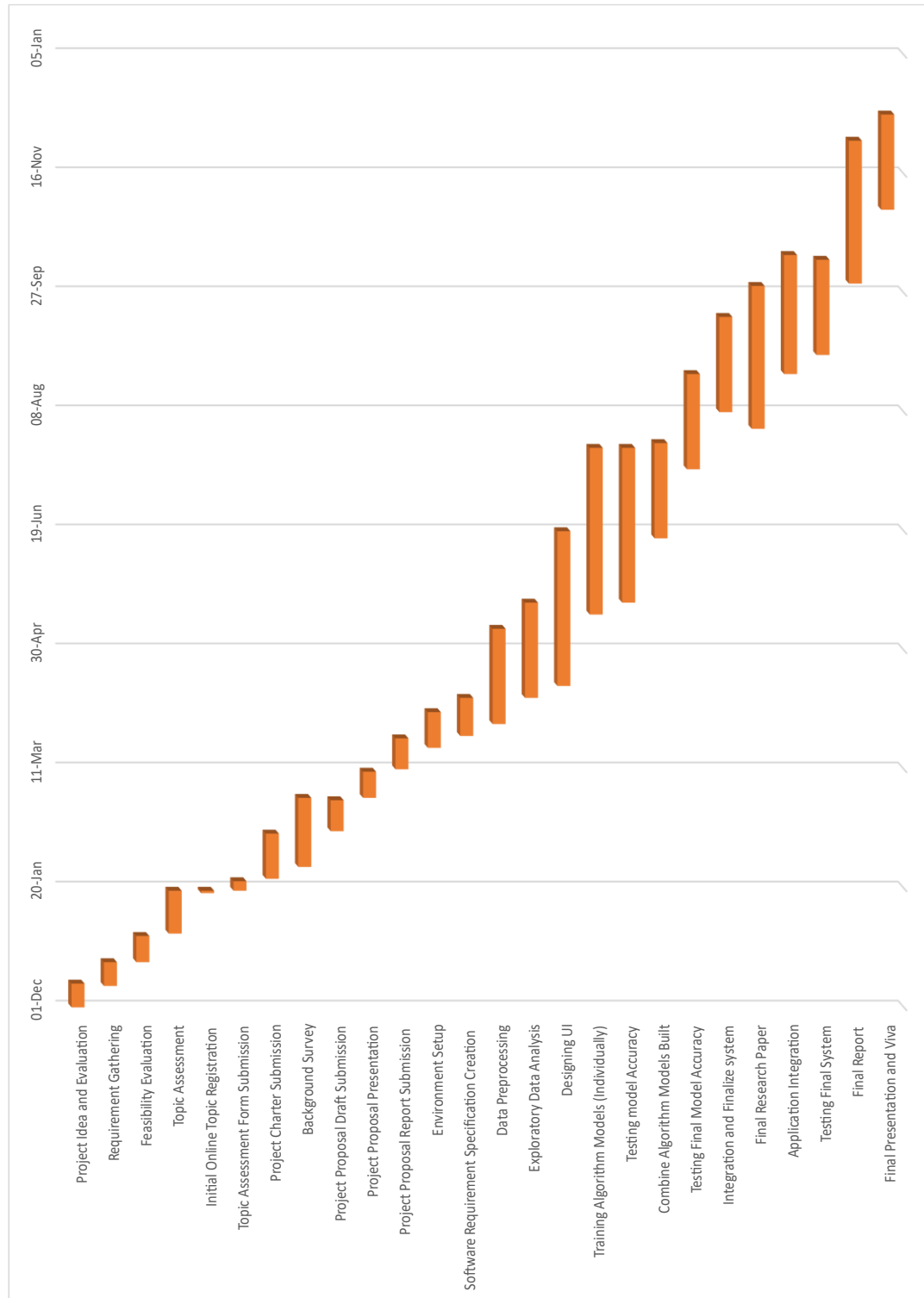


Figure 3.9 – Gantt Chart

3.7 Commercialization

The incorporation of a price prediction system within an online car classified portal can lead to higher and faster sales of the cars that are being advertised since the tagged price values are much more convenient, reliable, and realistic when comparing with other similar cars within the market and the buyers are more obliged to make a complete purchase consequently. Hence the business revenue of the application as a whole will be increased as well due to the higher number of customer visits and utilization. Customers would not need to do extensive research when clarifying a price value by exiting the application, which results in a decrease in the bounce rate. They would simply receive an approximated price value at the application venue itself and will be thoroughly satisfied with the service that saved their valuable time out of their busy schedules. Moreover, the inclusion of such a price prediction system within the application will give it a towering competitive advantage among other parallel systems that provide equivalent efforts in satisfying customer necessities. The summarization of the emphasized business value of the price prediction system's collaboration with the rest of the application is shown as below, in the Figure 3.8.



Figure 3.10 – Commercialization

3.8 Technologies To Be Used

- A Javascript front-end framework (Angular/React/Vue)
- Python
- Strapi (Express.js)
- Flask
- Tensorflow / Keras / Scikit-learn
- PyCharm / Webstorm / VSCode
- Docker
- Git
- MySQL / MongoDB
- Jenkins / Kubernetes

4 DESCRIPTION OF PERSONAL AND FACILITIES

Dynamic price prediction and suggestion

- Analyze price impactful item features and investigate for an efficient predictive algorithm model capable of forecasting a selling price for an item. (considering accuracy, efficiency, speed, effectiveness, size of training data, number of features productively supported, etc.)
- Develop and train the models' RF, SVM, and LightGBM, with refinements of parameters, using the preprocessed version of the 'Used Cars Dataset' acquired from Kaggle, having about 450,000 records and 26 car attributes.
- Build an ensemble model out of the above-stated three models and test for accuracy using a test dataset, to finalize the prediction engine.
- Build and develop necessary client-side and server-side components for data extraction of an item from advertisement form prior to posting.
- Integration with developed algorithm model and visualization of the predicted price for the customers in form of a suggestion.
- Continuous retraining of the model through a renewed dataset, extracted from current item information within the application itself since some customers submit their own price values, rather than the system suggest a price, or through unused records in the utilized dataset since those are necessary for future prediction processes to perform with continuous accuracy as well.

5 BUDGET AND BUDGET JUSTIFICATION

Table 5.1 – Budget

Component	Amount (USD)	Amount (LKR)
Domain Name Registration	12	2400
1GB Memory 1vCPU Droplet (Frontend)	60	12000
2GB Memory 1vCPU Droplet (Backend)	120	24000
Total	192	38400

REFERENCE LIST

- [1] NATIONAL TRANSPORT AUTHORITY. 2014. Available from: <http://nta.gov.mu/English/Statistics/Pages/Archive.aspx> [Accessed Feb. 19, 2021].
- [2] WOLF STREET. (2019). “Average Age of Vehicles Sets Record, New-Vehicle Sales Drop to Where They Were 20 Years Ago. What Are Automakers Doing?,” [online] Available at: <https://wolfstreet.com/2019/06/27/good-for-consumers-part-of-carmageddon-for-automakers-average-age-vehicles-in-operation/> [Accessed Feb. 19, 2021].
- [3] ACCESS. (2012). “Will China’s Vehicle Population Grow Even Faster than Forecasted?,” [Online] Available at: <https://www.accessmagazine.org/fall-2012/will-chinas-vehicle-population-grow-even-faster-forecasted/> [Accessed Feb. 19, 2021].
- [4] GREEN CAR CONGRESS. (2019). “US used vehicle sales are more than double the number of new vehicle sales” [Online] Available at: <https://www.greencarcongress.com/2019/07/20190716-fotw.html> [Accessed Feb. 19, 2021].
- [5] S. Pudaruth “Predicting the Price of Used Cars Using Machine Learning Techniques,” *International Journal of information & Computation Technology*, vol. 4, no. 7, pp. 753-764, 2014.
- [6] N. Kanwal and J. Sadaqat, “Vehicle Price Prediction System using Machine Learning Techniques,” *International Journal of Computer Applications*, vol. 167, no. 9, pp. 27–31, 2017.
- [7] S. Peerun, N. H. Chummun, and S. Pudaruth, “Predicting the Price of Second-hand Cars using Artificial Neural Networks,” *The Second International Conference on Data Mining, Internet Computing, and Big Data*, pp. 17–21, 2015.

- [8] E. Gegic, B. Isakovic, D. Keco, Z. Masetic, J. Kevric, “Car Price Prediction using Machine Learning Techniques,” *International Burch University, Sarajevo, Bosnia and Herzegovina*, vol. 8, no. 1, pp. 113-118, 2015.
- [9] N.Sun, H. Bai, Y. Geng, and H. Shi, “Price Evaluation Model in Second-hand Car System based on BP Neural Network Theory,” *18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pp. 431–436, 2017.
- [10] M. Listiani, “Support Vector Regression Analysis for Price Prediction in a Car Leasing Application,” *Master Thesis, Information and Media Technology, Hamburg University of Technology*, pp. 1-85, 2009.
- [11] N. Pal, P. Arora, S. Sumanth Palakurthy, D. Sundararaman, P Kholi, “How much is my car worth? A methodology for predicting used cars prices using Random Forest,” *Future of Information and Communications Conference (FICC)*, pp. 1-6, 2018.
- [12] Q. Truong, M. Nguyen, H. Dang, B. Mei, “Housing Price Prediction via Improved Machine Learning Techniques,” *International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI2019)*, pp. 433-442, 2020.
- [13] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *NIPS*.
- [14] Vidhi Chugh, “A Guide On When To Retrain Your Machine Learning Model,” towardsdatascience.com, Sep. 11, 2020. [Online]. Available: <https://towardsdatascience.com/when-are-you-planning-to-retrain-your-machine-learning-model-5349eb0c4706>. [Accessed Feb. 21, 2021].
- [15] Rasmus Steniche, “Why Retraining Is So Important,” neurospace.io, Sep. 27, 2019. [Online]. Available: <https://neurospace.io/blog/2019/09/why-is-retraining-so-important/> [Accessed Feb. 21, 2021].

- [16] Tin Kam Ho, "Random decision forests," *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, Canada, vol.1, pp. 278-282, 1995.
- [17] S. Russell, *Artificial Intelligence: A Modern Approach* (3rd edition). PE, 2015.
- [18] A. Ben-Hur, D. Horn, H.T. Siegelmann, V. Vapnik, "Support vector clustering," *Journal of machine learning research*, no. 2, pp. 125-137, 2001.
- [19] Aizerman, M. A. "Theoretical foundations of the potential function method in pattern recognition learning," *Automation and remote control*, no. 25, pp. 821- 837, 1964.
- [20] Necati Demir, "Ensemble Methods: Elegant Techniques to Produce Improved Machine Learning Products," toptal.com, 2016. [Online]. Available at: <https://www.toptal.com/machine-learning/ensemble-methods-machine-learning> [Accessed Feb. 22, 2021].
- [21] ADITI. (2005). "Agile Methodology based Services," [Online] Available at: <http://www.aditicorp.com/services/agile-methodology-based-services/> [Accessed Feb. 22, 2021].

APPENDIX

A: Plagiarism Report