Tievs: Classified Advertising Enhanced Using Machine Learning Techniques

Dhanuja Ranawake
Faculty of Computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
it18122060@my.sliit.lk

Savandi Bandaranayake
Faculty of Computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
it18113532@my.sliit.lk

Ravihari Jayasekara
Faculty of Computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
it18089400@my.sliit.lk

Imashi Madhushani
Faculty of Computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
it18082548@my.sliit.lk

Manori Gamage
Faculty of Computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
manori.g@sliit.lk

Suriyaa Kumari
Faculty of Computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
suriyaa.k@sliit.lk

***Abstract*— The scarce use of tangible periodicals led to a consistently soaring popularity of online classified advertising. Nevertheless, existing platforms retain complications. Most recommendation systems are built with conventional technologies that are less scalable, less accurate, and having high latency processes. Moreover, customers find it tiring when clarifying a reliable, precise price value for items they are trying to sell through the classified advertising system. Additionally, strict validation techniques to identify and prevent fraudulent content or images from being published in the advertising portals have been neglected. Therefore, authors have inaugurated a superior classified advertising system, Tievs, as a solution, by appraising said predicaments. It wields a flexible, process-simplifying, concurrency-induced recommendation breakthrough implemented from Universal Sentence Encoding incurred Natural Language Processing and Deep Learning routines. Furthermore, an innovative price prediction system having a supervised regression-based ensemble model forged ensuing a comparative study, having excellent accuracy in proactively predicting item prices as to cater hassles faced by customers, was satisfied. Light Gradient Boosting classifier-driven fake description analysis and a Convolution Neural Network powered figure deception recognition system were introduced, which gained prodigious precision with moral clarity in fraud detection and prevention. Hence, the proposed solution's objective of surpassing former classified advertising systems in delivering customers' necessities, using the most lucrative, time-saving, human-centric, and error-preventive approaches, was accomplished. It was affirmative by the positively responded questionnaire regulated among prospective users by the authors.**

Keywords*— *Classified advertising, ensemble learning, fraud detection, machine learning, optimized recommendation, price prediction

I. INTRODUCTION

Presently, the utilization of online platforms such as Craigslist, Olx, Quikr, Carewale, CarDheko, Facebook Marketplace, ikman.lk, etc., for classified advertisement interaction, has become increasingly popular compared to its predecessor periodicals, radio, and television communication channels. This inevitability, along with rapid technological growth and digitalization, has made it quite lucrative for both sellers and shoppers in their own endeavors whilst using the evolving marketplace of online classified ads [1]. Regardless of numerous implementations, a remarkably enhanced application that incorporates complex technologies to efficiently derive prominent functionalities that facilitate customer expectations and requirements has yet to arise. It is undoubtedly troublesome when interactions happen with an application that is not much accommodating to consumer needs appropriately. When recommending specific ads for customers with regard to their preferences, most systems do not account for meticulous characteristics of the descriptions within the searching products. Existing systems have considered common approaches for recommendation such as Term Frequency – Inverse Document Frequency (TF-IDF) rather than investigating more suitable innovative methods, i.e., Google Universal Sentence Encoder (USE) [2, 3]. Having less scalable, less accurate, high latency processes internally running using outdated common technologies degrades the overall system productiveness, causing a waste of time in finding the desired result, which eventually leads to high customer turnover rates as well.

Furthermore, it can be strenuous doing exhaustive research when clarifying the appropriate price for a product to be sold which reduces customer satisfaction and comfortableness. Usage of website crawlers within an application for target price realization is deemed idle due to their possible dependent server failures, incompatible or rigid attributes under consideration, failing requests from heavy network traffic, and future system augmentation immobilization. Hence an advertising platform providing its own independent product price prediction mechanism is more convenient and currently obscure. Moreover, from an application developer's perspective, they need to implement strict program logic to identify and prevent fraudulent advertisements from being submitted either by accident or deliberately, using verification and validation techniques. If those ads were submitted, in due course, shoppers

978-1-6654-0066-4/21/$31.00 ©2021 IEEE

may view erroneous ads and lose confidence in using the application for their needs prompting the downgrade of the platform's reputation. Current applications do not prioritize the above, let alone monitor and prevent them from occurring. Even more, many implemented classified advertising systems simply do not exhibit rich User Interfaces (UI s) for smooth functionality or promote quality user experience, in consonance with the latest trending scientific breakthroughs. After scrutinizing the prevailing issues, the authors have investigated and analyzed several scientific approaches and have proposed a novel solution as an intelligent and advanced web application that incorporates Machine Learning (ML) and Deep Learning (DL) technology, mainly aiming to surpass present-day resembling competitors and provide customers with the finest user experience when browsing online classifieds. The aforementioned aim was fulfilled by the achievement of an interactive and optimized recommendation system to serve ingenious methods to advocate preferred classifieds, a dynamic and independent price prediction procedure for specific products so that customers are saved from the burden of clarifying prices themselves, and the identification and prevention of fraudulent content and images before ad submission, leading to higher reliability and reputation of the application as a whole. The novelty aspect of the prospective classified ad portal is found through the combination of the above four components. This research was conducted by focusing primarily on **used car classifieds** to maintain scope.

The basic assembly of the research paper is as follows. Section II discusses similar background literature done previously on suggested components, while section III infers the methodology used in developing the proposed system within five sub-sections. Section IV interprets how the authors evaluated results, and lastly, section V demonstrates the conclusion with future work.

II. RELATED WORK

Due to inadequate existing research specifically covering the overall proposed solution, previous studies vaguely allying with the main four components of this research are being discussed.

Product recommendation employing sundry algorithms either by content-based, collaborative, or hybrid methods for movie, social networks, music, e-commerce systems etc., were targeted by studies done so far rather than heeding for classifieds portals. Authors R. Singla et al. [2] have done a content-based movie recommender engine and utilized 'Distributed Bag of Words' version of Paragraph Vector (PV-DBOW) and Term Frequency methods quite exceptionally for movie plots/descriptions examination and generation of an inter-movie similarity score while Badriyah et al. [3] developed a system to suggest properties by exerting content-based filtering powered with the TF-IDF method. Although admiring results have been obtained, [2]-[4] have overlooked an ideal technique that this paper bearded, which is USE, that acquire words and average them for the later transaction into a feedforward Deep Neural Network (DNN) to produce sentence embeddings quite efficiently [5]. Hence, the inferiority of PV-DBOW and TF-IDF techniques is comprehensible. Author S. Shaikh [6] used a graph-based approach for recommendation in an e-commerce website that utilizes an overlap semantic approach to integrate recommendation and semantics where overlap value is calculated using the specific formula and saved in a graph format. Only items with a correlation value of more than 0.4 were considered for recommendation. However, it is unusable for classified ad portals since user-preferred characteristics such as votes, likes/dislikes, ratings, views, feedback, or comments are not generally provided. Furthermore, it induces the Cold Start problem, where information is unavailable for the recommendation engine to handle newly introduced components efficiently. Therefore, our proposed system would not account for such attributes except for content-based features. Cosine-similarity fusion of these studies tempted this research's refinement in collaboration with the USE technique, which was not priorly investigated and exercised.

Supervised learning-based ML algorithms were commonly used for price prediction, either by classification or regression. Ganesh et al. [7] selected Multiple Regression, Lasso Regression, and Regression Trees algorithms to produce predictive models and used One-way Analysis Of Variance with Post-Hoc test to determine error rate variances which were quite avant-garde. Lasso and Regression Trees were deemed better, with error rates less than 5%. However, their results would have demeaned if a larger data sample was used to train and test than using only 804 records with 11 decisive features, being quite similar to Kuiper [8], who used a Multivariate regression model to predict numerical price values for 2005 General Motor cars using the same dataset extracted from pakwheels.com, encouraging the worth of variable selection techniques and analyzing multiple models' performances. Using the same dataset, Noor et al.[9] implied that Multiple Linear Regression might outperform all with its prediction precision of 98%, though it's unreliable since only 1699 records and 3 features were considered from feature engineering, reducing forecast perfection.

In [10], a comparative study examined 3 regression-based models where the Gradient Boosting model gave an admirable accuracy having mean absolute error as 0.28 in comparison to Random Forest (RF) and Multiple Linear Regression. Their staggering 304,133 data records with 11 predictors were applauding for model efficiency. Regardless, ensemble techniques yielding better performances were not explored in theirs, unlike this research. Pal et al. [11] investigated Linear Regression and RF, with 3 notable but ill-suited forecasters and RF triumphantly conferred train and test accuracies as 95.82% and 83.63%. Overfitting still seemed to linger in contemplation of the above values and despite cross-validation and grid search measures taken. One study [12] found RF classifier the best, having test accuracy of 83.08% with only 5 relevant features and withdrawing main ones as vehicle type, price, and transmission, surfacing a feature selection flaw as opposed to ours with 13 estimators. Authors [7]-[12] utilized datasets scraped from e-commerce websites and/or from Kaggle. Hence they originate the notion that peculiar car price knowledge is futile and confirms the sufficiency of open datasets that inspired this study as well. Contradictorily, Mauritius car data from newspapers were used by Pudaruth [13] to predict prices of used cars using Multiple Linear regression, Naïve Bayes, K-Nearest Neighbours, and Decision Tree algorithms. The classification accuracies were mediocre, i.e., below 70%, due to fewer car

observations at about 90 records, while the inability of Naïve Bayes and Decision Trees to predict continuous variables was indisputably perceptible. Assessments discussed [7]-[13] only analysed performance statistics of prediction for several acclaimed models individually. Divergently, Gegic et al. [14] affirmed the higher potential in accuracy and precision of an ensemble model's prediction, assembling Artificial Neural Network, Support Vector Machine, and RF classifiers having solitary correction less than 50%, and stating their combined accuracy on test data as 87.38%. Although having more samples than 797 could have improved prediction power and 90:10 split probably lead to model overfitting that reduced generalization levels, their articulation of the ensemble model is outstanding and encouraged this research to audition the same using regressors.

Sporadic attempts to distinguish and obviate online frauds have prevailed as Garg, and Nilizadeh [15] showed that vehicle frauds are prominent on craigslist in 30 locations in the United States (US). Approximately 1.7% of ads were detected as frauds incorporating indecorous keywords. Conversely, using a single indicator and having a limited sample size made it inconvenient to extrapolate their observations to other sorts of advertising. Another study [16] was the first to perform a statistical evaluation of 2 datasets to detect spam pages. They asserted that spammers utilize templates to create spam content automatically. As a result, even if specific words vary from one page to another, they will have an equal number of words. Two hundred sample-sized hosted web pages, including a minimum of 10 pages with no word count variation, were spammed 55%. Tran et al. [17] developed a classifier to detect spam on classified websites, and they contended that additional features such as occurrences of URLs or e-mail addresses are required to outperform traditional content or Natural Language Processing-based (NLP) classifiers. The abovementioned characteristics have incorporated domain-specific features such as the current market price of advertised items. They improved their F-1 scores by more than 50%. However, relying on external sources were hard to generalize characteristics which conveyed a significant drawback in their approach.

Two simple yet effective histogram and encoding-based detection methods were proposed by Yuanfang Guo et al. [18] for fake colorized images, which exhibit a proper execution against many state-of-the-art colorization approaches. Yet, their findings elucidate the declining effectiveness of their solution when the testing and training images are generated utilizing various colorization methods or datasets. Likewise, [19] introduced a novel rapid image matching technique, which constructs the k-d tree and uses better BBF algorithms to substitute the Linear algorithm and evaluate the performance of the amended method through testing. Consonantly, few web-based scam detection tools as Ads-Guard [20] assist novice users in ascertaining the legitimacy of certain ads, helping law enforcement inspections avert online classified ad crimes. Application Program Interfaces (APIs) and Image Recognition were used in [20] to estimates the likelihood of a particular classified being a scam by pasting its URL into the tool. Nevertheless, it had a few limits since it could only accept Craigslist advertisements. Fruitful precautions on fraud content/images have not been occupied within persisting classified advertising systems, thus precipitated this research to contrive a desirable measure to conceivably increase application confidentiality altogether.

III. Methodology

A. Training Data Source & Tools/Technologies Utilized

Two openly available Kaggle datasets were utilized for obligatory ML model rudiments. The 'Used Cars Dataset' sustaining 441,396 used car records covering about 40 brands having 25 unique attributes was scraped 2 months before this research inception from an American classified advertisement portal, namely 'Craigslist' since Tievs primitively targeted the US market. The 'Vehicle Detection Image Set' assisted the vehicle image recognition and fraud prevention strategy. Programming languages and frameworks as Python (Django, Flask/Quart), Angular (Typescript), Golang (Echo framework), and PHP were used with supporting libraries like Pandas, Keras, glob, NumPy, Scikit-learn, TensorFlow, Pyplot, Angular Material, etc. for ML model, front-end app, and model/Content Management System API developments. Jupyter Notebook, Anaconda Navigator, Visual Studio Code, PyCharm, WebStorm, and Nova Editor were used as technical tools while the database server amalgamated PostgreSQL. The Authors incorporated Postman to test REST APIs and the Supervisorctl system to revive APIs upon hosting failures.

B. Recommendation System

The following steps were exerted when implementing the ad feature assimilated recommendation engine, which is ingenious in nature and unfathomed within its field of research.

- Id and description columns of 18,035 car records were extracted after handling null and irrelevant values.
- Description content was cleansed with a complex regex function prior to lowercase conversion and **Penn Treebank tokenizing.**
- Then it's sent into **TensorFlow USE** trained with a **Deep Averaging Network (DAN)** encoder, which uses DNN to find the best combination of 512-dimensional vector embeddings is used upon the tokenized form to give NLP incurred ad relatedness. DAN trains the encoder itself by having preconfigured parameters built within its library for the encoder to map any text into embeddings, as shown in Figure 1 [21].

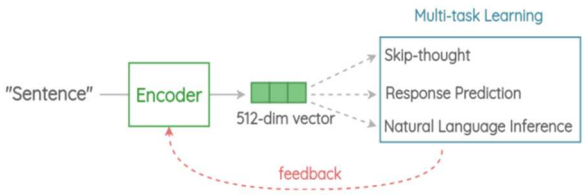


Fig. 1. USE embedding creation process [21]

- A **Golang API** was created to calculate the cosine similarity of those generated embeddings prior to sorting them from highest to lowest. New embedding generation for freshly submitted ads is supported from this API; hence concurrently handling the similarity calculations

promotes optimized CPU utilization and process efficiency.

- Finally, id and similarity scores are calculated and returned into the Tievs front-end app to be displayed according to customer preference.

Classifieds are often updated. On that account, imperative embedding creation when a new ad is submitted will not disturb the overall prompt, lower latency recommendation procedure. Thus, model retraining is deemed unnecessary, unlike other traditional recommendation techniques.

C. *Price Prediction System*

Exploratory data analysis (EDA) was done primarily to gain an expressive perception in the forms of bar charts, scatter plots, box plots, heat maps, distribution graphs, etc. and examine variable diversity, data field ranges, and correlation between each feature, including price. Correlation metrics showed that year and odometer were the most impactful for price. In view of a copious dataset and a limitation on computational resources, 100,000 records were accounted for price prediction model implementation. Below depicts some further data preprocessing steps undertaken in correspondence to regression-based ML algorithms.

- Omitting 11 advertisement attributes due to their proven insignificancy for model performance [22], and retaining impactful features as year, manufacturer, model, condition, cylinders, fuel, odometer, title status, transmission, drive, size, type, paint color (Feature engineering and selection).
- Filtering out records with unrealistic price values
- Removing records with null prices.
- Filling attribute null values with respective mean/mode values (Imputation).
- Selecting records where year >= 1975, price <= $100000 and price >= $750 (Handling Outliers).
- Categorical variable conversion to numerical with One-Hot-Encoding method which transforms values into either 1 or 0. It gives a realistic data interpretation than Label Encoding.

Authors conducted a comparative study focusing on the performance of 15 ML algorithms available in the Scikit-learn library, namely **Extra Trees, Bagging, RF, Gradient Boosting, Decision Trees, K-nearest Neighbors (KNN), Support Vector Machine (SVM), Lasso, Ridge, Elastic Net, AdaBoost, Multilayer Perceptron (MLP), Linear SVR, Stochastic Gradient Descent (SGD), and Multiple Linear Regressors**. Feature scaling (normalization) was used to deal with KNN, SVR, and SGD regressors. Each model was trained and evaluated with the same train, and test data associating **K-Fold Cross-Validation** with a split ratio of 75:25:10 multiple times to make the models less biased and account for every observation until satisfactory results emerge, and overfitting was eradicated to a commendable extent, enhancing generalization.

Regression-based model proficiency in forecasting continuous values [9], unselected classification route for this study. **RandomizedSearchCV and GridSearchCV** were used when determining the most optimized hyperparameters for the success of the above models. R-squared value (R2), adjusted R2, relative squared error (RSE), and root mean square error (RMSE) was used as accuracy criterion to determine the top 4 exemplary models. Those premium models were then ensembled together using a stacking regressor model as shown by Fig. 2, procuring an unprejudiced and sturdy prediction ability.

Algorithm Stacking

Input: training data $D = \{x_i, y_i\}_{i=1}^{m}$
Ouput: ensemble regressor H
Step 1: learn base-level regressor
for $t = 1$ to T do
  learn h_t based on D
end for
Step 2: construct new data set of predictions
for $i = 1$ to m do
  $D_h = \{x'_i, y_i\}$, where $x'_i = \{h_1(x_i), ..., h_T(x_i)\}$
end for
Step 3: learn a meta- regressor
learn H based on D_h
return H

Fig. 2. Stacking regressor pseudocode [22]

Ultimately, the model was serialized using Pickle and lodged with the Flask REST API's 'predict-price' endpoint for efficient price prediction and visualization through the advertisement submission form in Tievs front-end app.

D. *Fraud Content Detection System*

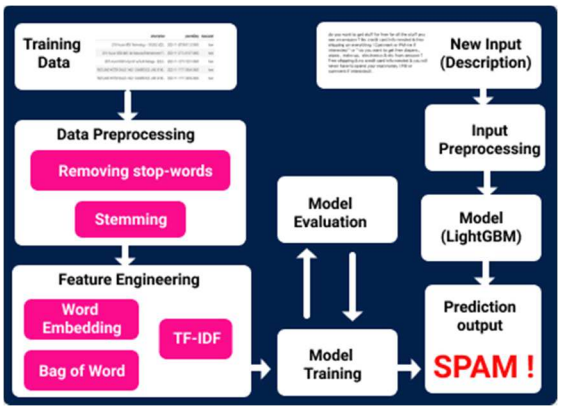


Fig. 3. Process diagram of the fraud detection model

Aiming to foresee the likelihood of fraudulent content prior to ad submission, an initial dataset of 23,000 records was labeled as legitimate and illegitimate, where the latter was prepared by the authors themselves. The fraud detection model involves a sequence of stages that constructs an effective ML model, as depicted in Fig 3. Succeeding EDA, **"Stop Words"** technique, which is often perceived in writing regardless of context, including words as "the," "a," "an," "but," "in," "because," etc. was utilized to eliminate all interference from inputs that would otherwise obstruct the potential to distinguish positive and negative classifiers. Snowball Stemmer algorithm, which eliminates affixes to retrieve the basic form of words, was utilized as a stemming technique due to its computational speed. Moreover, raw text was essentially converted into a number of vectors to perform in ML models due to its impotence to directly

operate textual forms. Then, a **Bag of Words (BOW)** model was developed to extract characteristics from the text before the implementation of the **TF-IDF NLP** vectorization technique to evaluate the significance of text data and serialize them for further transformation. Subsequently, the Scikit-learn library was utilized to convert text content into a matrix of numbers that could be fed to the classifier. This study was initiated to evaluate the expertise of 7 ML algorithms such as **Decision Tree, RF, SVM, KNN, XGBoost, XGBRF, and Light Gradient Boosting Machine (LightGBM) classifier** in a comparative analysis after vectorization. Each classifier was trained and tested in the same 75:25 ratio until the optimal model would be exposed in terms of F1 score (accuracy and recall) and confusion matrix. Based on iteration, the highest max depth value was identified from parameter tuning compared with depth-wise growth and competency to accomplish random grid selection of the most pertinent parameters for the best classifier. After successful embedding of the trained model in the REST API, it could classify ad content as legitimate or not and display caution Tievs near advertisement submission when indispensable.

E. Fraud Image Detection System

Authors stipulated the DL algorithm, **Convolution Neural Network (CNN),** and in its convolutional layer, initial image feature learning was done using the vehicle/non-vehicle image set to extract characteristics such as gradient orientation, colors, edges, etc., assign weights, and compress them dimensionally as convolved features compared to the original image. GREY scaling the data to execute image preparation procedures such as image rotation, resizing, and part extractions were done using Keras API. Next, the pooling layer reduced spatial size further while extracting superior traits and optimizing processing power. These 2 layers executed continuously and fastidiously trained the CNN model in discerning image qualities successfully.

The final output was flattened and column vectorized before classification through a fully connected feed-forward neural network layer. With backpropagation applied iterative training, the model could discriminate high/low-level image features preparatory to effectuating **Softmax classification**. After the model was thoroughly tested on the accuracy, it was capable of identifying ad permissible car images by the manufacturer as well as prohibited images. The specific ML API was created using the Django framework and integrated with the trained model before being implied into the front-end application to detect inappropriate images and display warning signs necessarily. CNN differentiates from the traditional model by using only a few parameters to reduce overfitting probability and considering neighbor context (images in this scenario) information to note similarities.

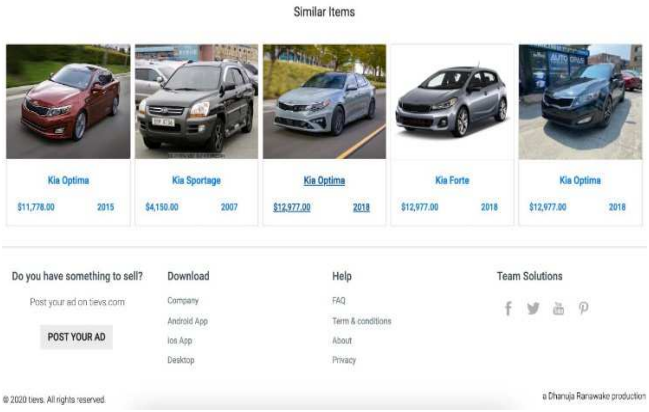
IV. Results & Discussion

The recommendation engine was simplified in implementation yet powerful in suggestion. It was found highly scalable, with an ability to support diverse scale systems, since when each advertisement is submitted, the specific embeddings created immediately contribute for cosine similarity calculation consequently, which is inclusively cost-efficient and resource optimizing. Golang implemented concurrency when calculating similarity enabled a swift response time of 6 to 10ms in a MacBook Air M1 8GB RAM when recommending equivalent ads. The below Figure. 4 illustrates Postman GET result arriving from the DB server itself (mentioned below) and its visualization in Tievs, relating to an inquired ad.

***** http://ml.tievs.com:8000/recommend/15043?category=vehicles**


```
{
    "id": 17004,
    "price": 40990,
    "year": 2019,
    "model": "camaro ss coupe 2d",
    "manufacturer": "chevrolet",
    "image_url": "https://images.craigslist.org/00h0h_3elhPqitghBz_0gw0co_600x450.jpg",
    "score": 0.9991744312412771
},
{
    "id": 17781,
    "price": 42990,
    "year": 2020,
    "model": "camaro ss coupe 2d",
    "manufacturer": "chevrolet",
    "image_url": "https://images.craigslist.org/00N0N_1xMPvfxRAIdz_0gw0co_600x450.jpg",
    "score": 0.9999822340568808
},
{
    "id": 17783,
    "price": 38990,
    "year": 2018,
    "model": "camaro ss coupe 2d",
    "manufacturer": "chevrolet",
```


(a)



(b)

Fig. 4. (a) Postman JSON result and (b) Tievs visualization of similar ads

The four most optimal vigorously functioning regressors for car price prediction of out 15 others were finalized by examining their respective performance indicators, as Table 1. depicts. They were later congregated together, forming a stacked heterogeneous ensemble model that makes powerful, robust predictions having lower error and variance rates without biasing singularly. Ensembled regressors having accuracy higher than 80% (comparing R2 and Adj. R2) signified eminent price evaluation power.

Price prediction is challenging altogether, and sufficient data cleaning plays a major role when impacting predictive performance, which this research could benefit more from, having a complex dataset to wield, as RMSE values reveal. Fig. 5 shows the mean score comparison between the stacking regressor and the individual models. Fig. 6 displays the predicted price range envision in the UI after unifying the developed ensemble model with the REST API and vital attribute accumulation from the advertisement form, essentially offering the liberty of selection for customers.

TABLE I. TRIUMPHANT MODELS QUALIFIED FOR ENSEMBLING

Price Prediction Model	R^2 (train)	R^2 (test)	Adj.R^2 (train)	Adj.R^2 (test)	RMSE (train)	RMSE (test)
Bagging Regressor	99.8%	89.6%	97.8%	86.7%	2102.5	4515.0
Extra Trees Regressor	98.0%	89.3%	99.8%	88.1%	577.2	4267.5
Random Forest Regressor	93.7%	87.9%	93.4%	84.5%	3492.2	4834.0
Gradient Boosting Regressor	86.7%	85.0%	86.0%	80.8%	5130.4	5376.4

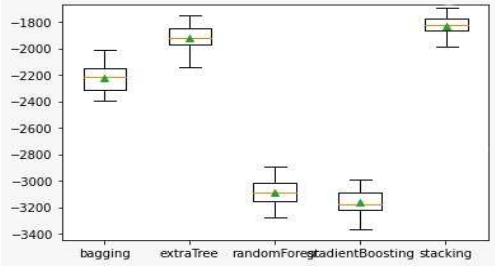


Fig. 5. Stacking regressor & individual model mean score comparison

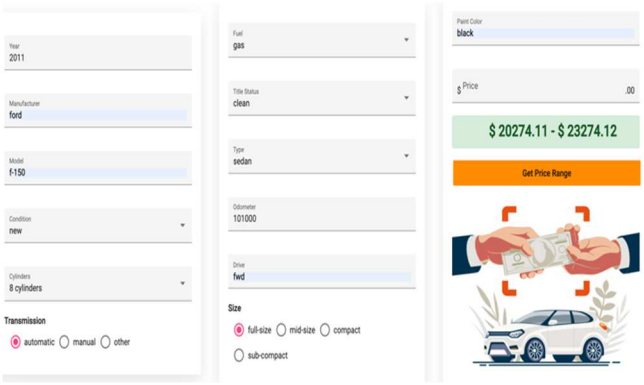


Fig. 6. Predicted price range visualization in front-end app

The heuristics utilized to detect and evaluate non-legitimate advertising were adequately precise. The non-legitimate made up 14.6% of the data, whereas legitimate composed 85.4% of the dataset. Although most classifiers appeared to be highly accurate, based on the F1 score, the LightGBM classifier was proven optimal, having 95.79% of precision, whereas the K-fold validation average accuracy was 97.8%. From iteration, max-depth of 5 had the highest F1 score of 0.8 according to random grid selection of the most pertinent parameters for the LGBM classifier. Pursuant to the confusion matrix shown in Fig. 7, the model properly classified 2200 records as legitimate and 4 records as fraudulent. Conversely, it has misidentified fraudulent content as legitimate content 0 times and 25 times misidentified legitimate content as fraudulent. An important aspect of this could be discerned in Table 2. The image recognition system was successfully able to detect 'fake' and 'not fake' images with minimal processing power and restrain inapposite images from being submitted into the Tievs application through deterrents. As given in Table 3., the CNN model accuracy rate is flattering, being above 90% with an adequate loss rate as well. Furthermore, underneath Fig. 6 renders the screenshot of the Tievs car classified advertising section which was constructed.

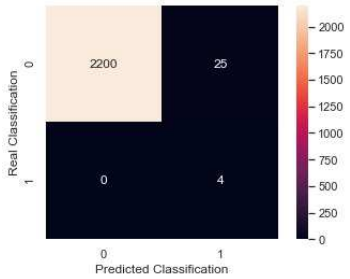


Fig. 7. Confusion matrix visualization

TABLE II. CLASSIFICATION REPORT OF LIGHTGBM CLASSIFER

Content Classification	F1 score	Recall	Precision	Support
Not Fraud	0.98	0.95	1.00	2200
Fraud	0.80	1.00	0.67	29

TABLE III. CONVOLUTION NEURAL NETWORK STATISTICS

Image Classification	Accuracy Rate	Loss Rate
Convolution Neural Network	93.4%	0.5031

Questionnaire Outcome for Tievs Application

The authors distributed a Google Form survey with 'Tievs' URL, which received 1005 responses, majorly from Males aging 20 to 60. Fig. 8 illustrates the portion of users who deemed the product in various satisfactory levels and Fig. 9 shows how productively each main feature achieved its respective goals.

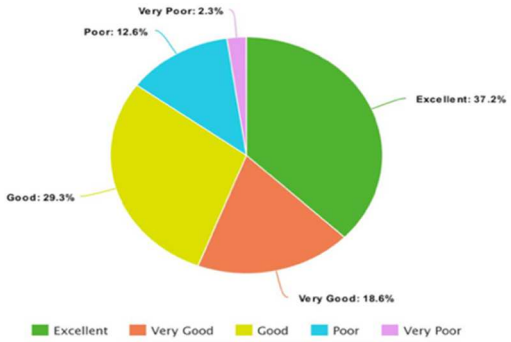


Fig. 8. Overall user review for system

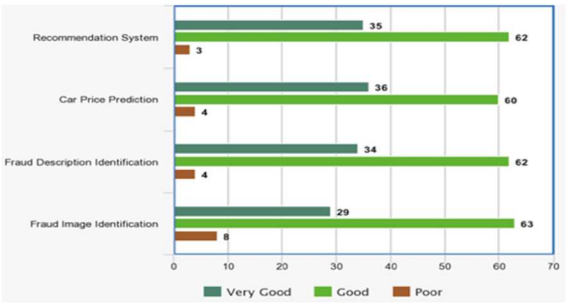


Fig. 9. Effectiveness of main components

V. Conclusion & Future Work

The novel recommendation engine on USE with DL and NLP techniques and cosine similarity calibrates the relatedness of classifieds according to user interacted ad details while reducing the cold start problem and making the process more simple and scalable. Cost-effectiveness made the application more trendy as opposed to traditional techniques. By collaborative filtering enhancements done hereafter where user data is acquired after application launch will provide more excellent classified advocacy. The regression-based ensemble model, consisting of 4 impenetrable regressors, provides strong price prediction for used cars with an accuracy of above 85% and even 90% in some instances, relieving the consumers from their intensive price value establishment when posting ads. Authors intend to tune model parameters while experimenting on more sophisticated genetic, fuzzy logic, and neural networks algorithms. Online model retraining is contemplated as well. Optimal classifier LightGBM achieved relevant parameters from analyzed models through supervised learning to identify fraud advertisement context before ad submission. Perhaps tweaking to the training phase with a larger fraud content dataset, presuming to augment recall rate precisely while reducing precision loss, could be done in the foreseeable future. Identifying fraudulent images was accomplished satisfactorily by giving attention to car image features using the CNN model, where it certified image legitimacy explicitly to prevent hoaxes. Furthermore, the authors anticipate determining the color of car images and their originality against submission as well. Nonetheless, this classified advertising web application was heightened with beneficial qualities to emphasize online advertising stature and score competitive advantages against other prevailing systems. As a future avenue for the overall design, authors intend to expand the services presently orchestrated by embracing different classified types such as real estate, electronics, pets, jobs, etc. Hence broadening the target consumer audience altogether would be plausible.

Acknowledgment

The authors highly appreciate the continuous guidance and encouragement provided by Sri Lanka Institute of Information Technology lecturers, colleagues, and family members. Special thanks are given to Mr. Austin Reese, Mr. Baris Dincer, and Kaggle team for developing and making analysis-foundational datasets openly available for all researchers and tech enthusiasts.

References

[1] J. Meffert, D. Morawiak, T. Schumacher, “Online classified ads: Digital, dynamic, and still evolving”, Telecommunication, Media, Technology, McKinsey&Company, [online document], 2015. Available at: http://www.mckinsey.com [Accessed: July 20, 2021].

[2] R. Singla, S. Gupta, A. Gupta, and D. K. Vishwakarma, “FLEX: A Content Based Movie Recommender,” in 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1–4, doi: 10.1109/INCET49848.2020.9154163.

[3] T. Badriyah, S. Azvy, W. Yuwono and I. Syarif, "Recommendation system for property search using content based filtering method," 2018 International Conference on Information and Communications Technology (ICOIACT), 2018, pp. 25-29, doi: 10.1109/ICOIACT.2018.8350801.

[4] A. Pal, P. Parhi, and M. Aggarwal, “An improved content based collaborative filtering algorithm for movie recommendations,” in 2017 Tenth International Conference on Contemporary Computing (IC3), 2017, pp. 1–3, doi: 10.1109/IC3.2017.8284357.

[5] Cer, Daniel & Yang, Yinfei & Kong, Sheng-yi & Hua, Nan & Limtiaco, Nicole & John, Rhomni & Constant, Noah & Guajardo-Cespedes, Mario & Yuan, Steve & Tar, Chris & Sung, Yun-Hsuan & Strope, Brian & Kurzweil, Ray, “Universal Sentence Encoder”, 2018.

[6] S. Shaikh, S. Rathi and P. Janrao, "Recommendation System in E-Commerce Websites: A Graph Based Approached," 2017 IEEE 7th International Advance Computing Conference (IACC), Hyderabad, India, 2017, pp. 931-934, doi: 10.1109/IACC.2017.0189.

[7] Ganesh, Mukkesh & Venkatasubbu, Pattabiraman, “Used Cars Price Prediction using Supervised Learning Techniques”, *International Journal of Engineering and Advanced Technology*, vol. 9, Dec., pp. 216-223, 2019.

[8] Kuiper, Shonda. "Introduction to Multiple Regression: How Much Is Your Car Worth?", *Journal of Statistics Education 16.3*, 2008.

[9] N. Kanwal and J. Sadaqat, “Vehicle Price Prediction System using Machine Learning Techniques,” *International Jounal of Computer Applications,* vol. 167, no. 9, pp. 27–31, 2017.

[10] N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya and P. Boonpou, "Prediction of prices for used car by using regression models," 2018 5th International Conference on Business and Industrial Research (ICBIR), 2018, pp. 115-119, doi: 10.1109/ICBIR.2018.8391177.

[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), 2018, pp. 1-6.

[12] Xinyuan Zhang , Zhiye Zhang and Changtong Qiu, “Model of Predicting the Price Range of Used Car”, 2017.

[13] 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.

[14] 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.

[15] V. Garg and S. Nilizadeh, “Craigslist Scams and Community Composition: Investigating Online Fraud Victimization,” in *2013 IEEE Security and Privacy Workshops*, San Francisco, CA, USA, May 2013, pp. 123–126. doi: 10.1109/SPW.2013.21.

[16] D. Fetterly, M. Manasse, and M. Najork, “Spam, damn spam, and statistics: using statistical analysis to locate spam web pages,” in Proceedings of the 7th International Workshop on the Web and Databases colocated with ACM SIGMOD/PODS 2004 – WebDB 04, Paris, France, 2004, p. 1. doi: 10.1145/1017074.1017077.

[17] H. Tran, T. Hornbeck, V. Ha-Thuc, J. Cremer, and P. Srinivasan, “Spam detection in online classified advertisements,” ACM Int. Conf. Proceeding Ser., vol. 11, Apr. 2011, doi: 10.1145/1964114.1964122.

[18] Y. Guo, X. Cao, W. Zhang, and R. Wang, “Fake Colorized Image Detection,” *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 8, pp. 1932–1944, Aug. 2018, doi: 10.1109/TIFS.2018.2806926.

[19] J. Xingteng, W. Xuan, and D. Zhe, “Image matching method based on improved SURF algorithm,” in 2015 IEEE International Conference on Computer and Communications (ICCC), Oct. 2015, pp. 142–145. doi: 10.1109/CompComm.2015.7387556.

[20] S. Al-Rousan, A. Abuhussein, F. Alsubaei, L. Collen, and S. Shiva, “Ads-Guard: Detecting Scammers in Online Classified Ads,” in *2020* IEEE Symposium Series on Computational Intelligence (SSCI*)*, Dec. 2020, pp. 1492–1498. doi: 10.1109/SSCI47803.2020.9308544.

[21] A. Chaudhary, “Universal Sentence Encoder visually explained,” *Amitness.com*, 15-Jun-2020. [Online]. Available: https://amitness.com/2020/06/universal-sentence-encoder/. [Accessed: July 24, 2021].

[22] KDnuggets, “Ensemble Learning to Improve Machine Learning Results - KDnuggets,” KDnuggets, 2021. [Online]. Available: https://www.kdnuggets.com/2017/09/ensemble-learning-improve-machine-learning-results.html/2 [Accessed: July. 25, 2021]