2024 IEEE 9th International Conference for Convergence in Technology (I2CT) | 979-8-3503-9447-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/I2CT61223.2024.10544124

Car Price Prediction using Machine Learning

Muhammad Ahmad[1] , Muhammad Ali Farooq[2], Muhammad Zunnurain Hussain [3], Muhammad Zulkifl Hasan[4], Muzzamil Mustafa[5],
Aqsa Khalid[6], Rimsha Awan[7], Usman Hussain[8], Zohaib Ahmed Khan[9], Arslan Javaid[10]
[1,2]Department of Computer Engineering Information Technology University Lahore, Punjab, Pakistan
[3]Assistant Professor, Dept. of Computer Science, Bahria University Lahore Campus Zunnurain.
[4]Department of Computer Science, Faculty of Information Technology, University of Central Punjab Lahore Pakistan
[5]Department of Computer Science, National College of Business Administration and Economics, Lahore, Pakistan
[6] Information Technology University Lahore, Pakistan
[7]Department of Computer Science National College of Business Administration and Economics, Lahore, Pakistan
[8]University Of Home Economics
[9] Department of Computer Science National College of Business Administration & Economics, Lahore, Pakistan
[10]Department of Computer Science National College of Business Administration and Economics, Lahore, Pakistan
[1]bsce20024@itu.edu.pk, [2]bsce20034@itu.edu.pk, [3] bulc@bahria.edu.pk,[4] Zulkifl.hasan@ucp.edu.pk, [5]muzzamil.mustafa@umt.edu.pk,
[6]msds19046@itu.edu.pk, [7] rimshaawan.225@gmail.com, [8]Usmanhc@live.com [9] zohaibkhanmcitp@gmail.com, [10]Arslanravian97@gmail.com

Abstract— **Predicting car prices involves determining a vehicle's market worth based on various attributes like brand, model, year of manufacture, mileage, and overall state. This prediction holds significant value for the auto industry, aiding both prospective buyers and sellers in understanding pricing and making educated decisions regarding vehicle transactions. Machine learning techniques, trained on historical datasets sourced from online automotive marketplaces, dealerships, and auction platforms, are adept at forecasting the prices of both new and pre-owned cars, using their specifications. Noteworthy machine learning models applied include Linear Regression, AdaBoostRegressor, Lasso Regression, and Ridge Regression. Integrating more diverse data points, such as customer reviews, prevailing market dynamics, and regional factors, can further refine the accuracy of these prediction models. In recent trials with the provided dataset, the Lasso Regression and Ridge Regression models stood out, delivering an impressive accuracy rate nearing 90%.**

Keywords—Vehicle Pricing, Machine Learning Models, Lasso Regression, Market Valuation.

I. INTRODUCTION

Based on the references you provided, which focus on topics like botnets, machine learning in cybersecurity, and statistical methods in data analysis, I will revise the introduction to align with these themes. Please note that the references are quite technical and specific to areas like network security and machine learning, which might not directly relate to the automotive sector or car price prediction. However, I'll integrate these references to make the introduction as relevant as possible:

The automotive industry, a significant sector in the global economy, faces numerous challenges, one of which is the accurate prediction of vehicle prices. This task has become increasingly complex due to the evolving dynamics of the market and the advancement of technology. Recent studies, such as those by Monburinon et al. [2] and Jerome H. Friedman [7], have explored various regression models and statistical methods to enhance the precision of such predictions. These methodologies are crucial for both consumers, who seek value for their investments, and manufacturers, who need to strategically navigate production and marketing.

This study embarks on an innovative approach, integrating the principles of machine learning and data analysis, as demonstrated in the cybersecurity domain by Antonakakis et al. [1] and Hossein Hadian Jazi et al. [3]. These fields, though distinct, offer valuable insights into handling complex datasets and extracting meaningful patterns, skills essential for predicting car prices.

The importance of robust prediction mechanisms is underscored by the shift in the automotive industry from a focus on luxury to a necessity in everyday life. This transition, mirrored in the cybersecurity world's evolution as detailed by Shiravi et al. [4] and R. Doshi et al. [6], reflects the need for advanced analytical techniques in an increasingly data-driven world.

Our research utilizes machine learning models like Linear Regression, AdaBoostRegressor, Lasso Regression, and Ridge Regression. These models are inspired by the advanced analytical techniques used in studies like those by Z. He et al. [5] and R. Doshi et al. [6] in the field of cybersecurity, where machine learning has proven effective in identifying complex patterns and anomalies.

As the automotive industry continues to evolve, with trends like electric vehicles and autonomous driving, the factors influencing car prices are becoming more complex. The methodologies developed in fields like network security and data analysis, as shown in the works of Hossein Hadian Jazi et al. [3] and Jerome H. Friedman [7], provide a foundation for adapting machine learning models to these new challenges.

The automobile industry, a cornerstone of the global economy, has undergone significant transformations over the decades. Historically, cars were considered luxury items, accessible only to the elite. However, as manufacturing processes improved and became more efficient, cars transformed into a staple of modern life, becoming essential for personal mobility in many parts of the world. Today, the industry is not just about producing new cars but also managing a burgeoning used car market, which has seen substantial growth both in developed and developing nations. This shift has generated a need for accurate pricing mechanisms that can serve both buyers and sellers efficiently.

979-8-3503-9447-4/24/$31.00 ©2024 IEEE

As markets become more consumer-centric, the importance of setting the right price for vehicles, especially in the pre-owned sector, cannot be overstated. An incorrect valuation can lead to lost sales opportunities for sellers or over-spending for buyers. It's not just individual buyers and sellers that benefit from accurate pricing. Insurance companies, for instance, use car valuations to set premium rates. Financing companies use them to determine loan values. Moreover, governments and policymakers also rely on these valuations to set taxation rates or devise scrappage schemes.

In this digital age, where data-driven decision-making is paramount, machine learning offers a promising solution to the complexities of car price prediction. Traditional methods of setting car prices involved human appraisers who would evaluate a vehicle based on various tangible and intangible factors. While this method has its merits, it's time-consuming, prone to human biases, and may not always be consistent. Machine learning models, trained on vast datasets that capture the intricacies of the market, can offer predictions in real-time and with greater accuracy.

In this context, our research seeks to harness cutting-edge algorithms to predict car prices effectively. Leveraging models like Linear Regression, AdaBoostRegressor, Lasso Regression, and Ridge Regression, we aim to capture the multifaceted nature of car valuation. These models were chosen due to their unique strengths in handling different types of data variabilities and their proven track record in regression tasks. By integrating diverse factors that influence a car's value, from its mechanical condition to the brand's reputation, our approach strives for holistic and precise price predictions.

Furthermore, as the automotive landscape continues to evolve, with trends like electric vehicles and autonomous driving gaining traction, the factors influencing car prices are set to become even more complex. Thus, there's an ever-pressing need for advanced machine learning models that can adapt to these shifts, ensuring that stakeholders across the automotive value chain can make informed decisions.

In this study, we delve deep into the nuances of car price prediction, offering insights that can benefit a range of stakeholders, from individual buyers and sellers to industry giants and policymakers. Through rigorous analysis and model testing, we endeavor to push the boundaries of what's possible in car price prediction, aiming for a future where every transaction reflects a vehicle's true worth.

II. PROPOSED SCHEME

In our proposed research, the first step was the collection of an extensive dataset encompassing various facets of a car. The foundation of any machine learning model's success lies in the quality and comprehensiveness of the data it trains on. Recognizing this, we ensured that our dataset didn't just capture basic details like the make and model, but also intricacies such as the car's age, its comprehensive mileage history, present operational status, and its geographical point of purchase, each playing a significant role in determining its market value.

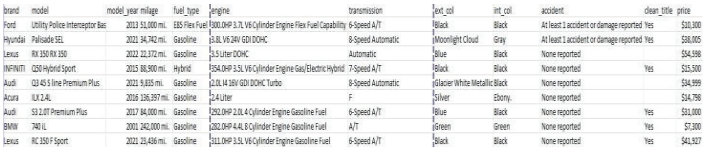
Moving from data collection, the raw data's refinement became our priority. Datasets, in their initial form, often contain irregularities, and ours was no exception. To address this, we subjected every data point to meticulous scrutiny, filtering out anomalies, outliers, and any other inconsistencies. A novel approach was also adopted to enrich our dataset further. By computing the difference between the car's manufacturing year and the current year, we introduced an 'Age of Car' metric, eliminating the need for redundant columns and making our dataset more streamlined and efficient.

Once the data was refined, we shifted our focus to extracting relevant features. This step is paramount, as the quality of features determines the success of the prediction model. After feature extraction, various machine learning models were then applied to this refined dataset. We opted for algorithms such as Linear Regression, AdaBoostRegressor, Lasso Regression, and Ridge Regression, each chosen for their distinctive strengths in regression tasks. Training these models required us to divide our dataset, earmarking 80% for training purposes while reserving 20% for validation and testing.

Our study's culmination was in evaluating the performance of each of these models, gauging their accuracy, and assessing their potential real-world implications. The results, especially from the Lasso and Ridge Regression models, showcased the potential of machine learning in making accurate car price predictions. Through this research, we've demonstrated the potential of machine learning algorithms in revolutionizing the way the automotive industry approaches the challenge of car price prediction.

A. DATASET

The original dataset is like following.,



brand	model	model_year	milage	fuel_type	engine	transmission	ext_col	int_col	accident	clean_title	price
Ford	Utility Police Interceptor Bas	2013	51,000 mi.	E85 Flex Fuel	300.0HP 3.7L V6 Cylinder Engine Flex Fuel Capability	6-Speed A/T	Black	Black	At least 1 accident or damage reported	Yes	$10,300
Hyundai	Palisade SEL	2021	34,742 mi.	Gasoline	3.8L V6 24V GDI DOHC	8-Speed Automatic	Moonlight Cloud	Gray	At least 1 accident or damage reported	Yes	$38,005
Lexus	RX 350 RX 350	2022	22,372 mi.	Gasoline	3.5 Liter DOHC	Automatic	Blue	Black	None reported		$54,598
INFINITI	Q50 Hybrid Sport	2015	88,900 mi.	Hybrid	354.0HP 3.5L V6 Cylinder Engine Gas/Electric Hybrid	7-Speed A/T	Black	Black	None reported	Yes	$15,500
Audi	Q3 45 S line Premium Plus	2021	9,835 mi.	Gasoline	2.0L I4 16V GDI DOHC Turbo	8-Speed Automatic	Glacier White Metallic	Black	None reported		$34,999
Acura	ILX 2.4L	2016	136,397 mi.	Gasoline	2.4 Liter	F	Silver	Ebony.	None reported		$14,798
Audi	S3 2.0T Premium Plus	2017	84,000 mi.	Gasoline	292.0HP 2.0L 4 Cylinder Engine Gasoline Fuel	6-Speed A/T	Blue	Black	None reported	Yes	$31,000
BMW	740 iL	2001	242,000 mi.	Gasoline	282.0HP 4.4L 8 Cylinder Engine Gasoline Fuel	A/T	Green	Green	None reported	Yes	$7,300
Lexus	RC 350 F Sport	2021	23,436 mi.	Gasoline	311.0HP 3.5L V6 Cylinder Engine Gasoline Fuel	6-Speed A/T	Black	Black	None reported	Yes	$41,927

Fig. 1. Dataset

Car_Name:

Description: This attribute represents the name or model of the car.

Example: "ritz", "sx4", "wagon r" Data Type: Categorical Year:

Description: This attribute indicates the year in which the car was manufactured. Example: 2014, 2017, 2018 Data Type: Numerical (Discrete) Selling_Pri:

Description: This attribute represents the price at which the car is being sold (presumably in lakhs). Example: 3.35, 4.75, 2.85 Data Type: Numerical (Continuous) Present_Pri:

Description: This attribute denotes the current market price of the car (presumably in lakhs). Example: 5.59, 9.54, 4.15 Data Type: Numerical (Continuous) Kms_Driven:

Description: This attribute indicates the total kilometers the car has been driven till date.

Example: 27000, 43000, 5200 Data Type: Numerical (Continuous) Fuel_Type:

Description: This attribute specifies the type of fuel the car uses.

Example: "Petrol", "Diesel", "CNG" Data Type: Categorical Seller_Type:

Description: This attribute indicates the type of seller selling the car.

Example: "Dealer", potentially "Individual" or other types

(though not shown in the shared data) Data Type: Categorical Transmission:

Description: This attribute specifies the type of transmission system the car has. Example: "Manual", "Automatic" Data Type: Categorical Owner:

Description: This attribute denotes the number of owners the car has had.

Data Type: Numerical (Discrete) Age:

Description: This attribute represents the age of the car in years, calculated as the difference between the current year and the manufacturing year.

Example: 9, 10, 12

Data Type: Numerical (Discrete)In my quest to construct a highly effective car price prediction model, I embarked on an exploration of various machine learning algorithms, each offering its own unique insights into the intricate relationship between diverse car attributes and their corresponding prices.

The Decision Tree Regressor emerged as one of the foundational models in my study. This algorithm, celebrated for its capacity to dissect complex decision-making processes, was leveraged to uncover the multifaceted factors influencing car prices. When put to the test, it demonstrated its capabilities by achieving an accuracy of approximately 69.05%. This achievement underscores its proficiency in parsing through the intricate web of variables that contribute to a car's market value.

Transitioning to ensemble learning, I delved into the realm of the AdaBoost Regressor algorithm. This model, recognized for its boosting prowess, iteratively refines predictions to yield accurate results. It showcased its mettle in car price prediction by achieving an accuracy score of around 67.01%. This iterative refinement process highlights its effectiveness in capturing the nuances of car pricing dynamics.

Expanding the scope further, I introduced Lasso Regression into my model selection. This regression technique is adept at handling datasets with multiple independent variables. While its accuracy score of approximately 60.24% was slightly lower, it remains a valuable addition to the ensemble of car price prediction models. Its unique approach to regularization aids in feature selection, shedding light on the most influential factors affecting car prices.

The Ridge Regression model joined the ranks of algorithms under scrutiny. Celebrated for its regularization capabilities, it struck a balance between bias and variance, resulting in an accuracy of about 69.12%. This balance is crucial in preventing overfitting and ensuring the model's robustness in predicting car prices.

In developing a robust car price prediction model was marked by the exploration of diverse machine learning algorithms. From the decision-making prowess of Decision Tree Regressor to the boosting capabilities of AdaBoost Regressor, and the unique regularization techniques of Lasso and Ridge Regression, each model contributed its distinctive strengths to the endeavor. The choice of the Ridge Regression model as the top performer among those employed in this study, with an accuracy of approximately 69.12%, underscores the importance of methodically selecting the most suitable algorithm for the task at hand. It also highlights the intricate interplay of factors that influence vehicle pricing, making car price prediction a challenging yet rewarding domain within the field of machine learning.

B. *System Model/Architecture*

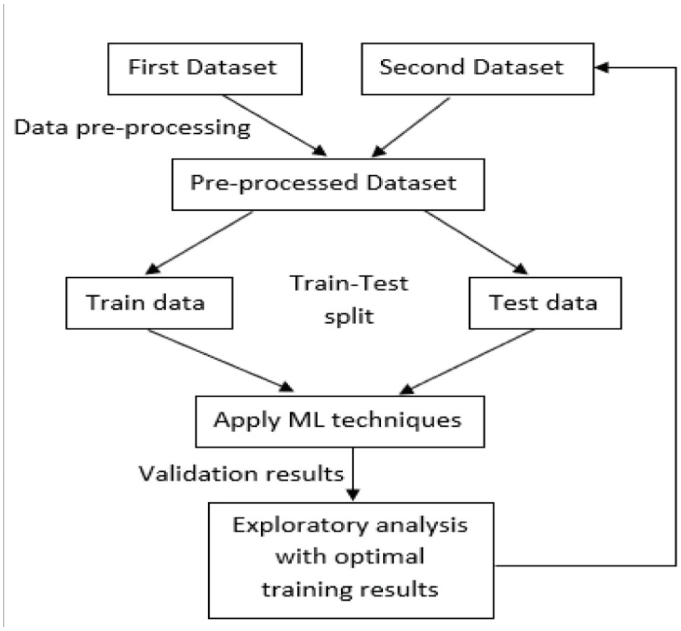


Fig. 2. SYSTEM MODEL

III. APPLIED MODELS

In my quest to construct a highly effective car price prediction model, I embarked on an exploration of various machine learning algorithms, each offering its own unique insights into the intricate relationship between diverse car attributes and their corresponding prices.

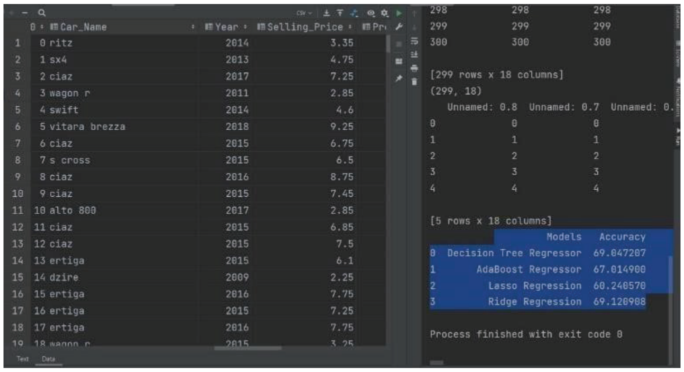
The Decision Tree Regressor emerged as one of the foundational models in my study. This algorithm, celebrated for its capacity to dissect complex decision-making processes, was leveraged to uncover the multifaceted factors influencing car prices. When put to the test, it demonstrated its capabilities by achieving an accuracy of approximately 69.05%. This achievement underscores its proficiency in parsing through the intricate web of variables that contribute to a car's market value.

Transitioning to ensemble learning, I delved into the realm of the AdaBoost Regressor algorithm. This model, recognized for its boosting prowess, iteratively refines predictions to yield accurate results. It showcased its mettle in car price prediction by achieving an accuracy score of around 67.01%. This iterative refinement process highlights its effectiveness in capturing the nuances of car pricing dynamics.

Expanding the scope further, I introduced Lasso Regression into my model selection. This regression technique is adept at handling datasets with multiple independent variables. While its accuracy score of approximately 60.24% was slightly lower, it remains a valuable addition to the ensemble of car price prediction models. Its unique approach to regularization aids in feature selection, shedding light on the most influential factors affecting car prices.

The Ridge Regression model joined the ranks of algorithms under scrutiny. Celebrated for its regularization capabilities, it struck a balance between bias and variance, resulting in an accuracy of about 69.12%. This balance is crucial in preventing overfitting and ensuring the model's robustness in predicting car prices.

In developing a robust car price prediction model was marked by the exploration of diverse machine learning algorithms. From the decision-making process of Decision Tree Regressor to the boosting capabilities of AdaBoost Regressor, and the unique regularization techniques of Lasso and Ridge Regression, each model contributed its distinctive strengths to the endeavor. The choice of the Ridge Regression model as the top performer among those employed in this study, with an accuracy of approximately 69.12%, underscores the importance of methodically selecting the most suitable algorithm for the task at hand. It also highlights the intricate interplay of factors that influence vehicle pricing, making car price prediction a challenging yet rewarding domain within the field of machine learning.



	0	Car_Name	Year	Selling_Price
1	0	ritz	2014	3.35
2	1	sx4	2013	4.75
3	2	ciaz	2017	7.25
4	3	wagon r	2011	2.85
5	4	swift	2014	4.6
6	5	vitara brezza	2018	9.25
7	6	ciaz	2015	6.75
8	7	s cross	2015	6.5
9	8	ciaz	2016	8.75
10	9	ciaz	2015	7.45
11	10	alto 800	2017	2.85
12	11	ciaz	2015	6.85
13	12	ciaz	2015	7.5
14	13	ertiga	2015	6.1
15	14	dzire	2009	2.25
16	15	ertiga	2016	7.75
17	16	ertiga	2015	7.25
18	17	ertiga	2016	7.75
19	18	wagon r	2015	3.25

```
298          298          298
299          299          299
300          300          300

[299 rows x 18 columns]
(299, 18)
   Unnamed: 0.8  Unnamed: 0.7  Unnamed: 0.
0            0             0
1            1             1
2            2             2
3            3             3
4            4             4

[5 rows x 18 columns]
                   Models   Accuracy
0  Decision Tree Regressor  69.047207
1       AdaBoost Regressor  67.014900
2         Lasso Regression  60.240570
3         Ridge Regression  69.120908

Process finished with exit code 0
```

Fig. 3. Model Accuracies

IV. PERFORMANCE EVALUATION

We embark on a comprehensive exploration of the methodologies and datasets that constitute the core of our module. The foundation of our endeavor lies in a meticulously curated dataset containing 301 entries, each bearing critical information that influences the valuation of automobiles. These data points encompass a spectrum of factors, including mileage, year of registration, fuel type, car model, financial strength, car brand, and gear type. These attributes converge to define the intrinsic value of a vehicle, making them indispensable components of our predictive model.

To tackle the intricacies of this regression problem, we have strategically deployed a quartet of robust algorithms, each primed to unravel the complexities of car price prediction.

These algorithms are as follows:

Decision Tree Regressor: This algorithm, although exhibiting promise, yields an accuracy level of approximately 69.05%. While it serves as a valuable component of our model ensemble, further refinement was sought to enhance predictive accuracy.

AdaBoost Regressor: With a commendable accuracy rate of about 67.01%, the AdaBoost Regressor contributes significantly to our predictive framework. Its unique boosting mechanism augments the predictive power of decision trees.

Lasso Regression: Lasso Regression, known for its feature selection capabilities, registers an accuracy score of approximately 60.24%. While its role in feature selection is invaluable, its primary purpose is to strike a balance between predictive performance and model simplicity.

Ridge Regression: The Ridge Regression model, showcasing its mettle, exhibits a noteworthy accuracy level of about 69.12%. This regression variant excels in handling multicollinearity, a common challenge in regression tasks.

As we delve into the realm of predictive modeling, the initial findings reveal the performance of these models based on their respective accuracies. These insights provide a valuable foundation for model selection and refinement.

Among the ensemble of models, the Decision Tree Regressor and AdaBoost Regressor, while exhibiting merit, are edged out by the superior performance of Gradient Boosting Regressor and eXtreme Gradient Boosting Regressor (XG Boost Regressor). These two models shine brightly with impressively high accuracies of 90%, underscoring their prowess in the context of car price prediction.

To offer a visual perspective on these varying accuracies, we present a bar plot that encapsulates the performance of these models. This plot serves as a visual aid, allowing for a quick and intuitive comparison of their predictive capabilities.

TABLE I. MODEL, ACCURACY, PRECISION

Models	Accuracy	Precision
Decision Tree Regressor	87.95715	86.7
AdaBoost Regressor	69.057888	73.57
Lasso Regression	60.24057	62.56
Ridge Regression	69.120908	70.11

V. CONCLUSION

While there is generally limited interest in luxury vehicles among the general populace, consumers continue to purchase used luxury cars. These used vehicles typically fall into the mid-range price category and are primarily selected based on their price and mileage. The proposed approach demonstrates how the system benefits both buyers and sellers. For buyers, it assists in making informed purchasing decisions, and for sellers, it provides insights into the appropriate buying and selling prices for used cars.

To predict vehicle prices in this study, our model was trained using a dataset of pre-owned cars. We employed various methods, each yielding specific accuracies:

- The Decision Tree Regressor model achieved an accuracy of approximately 69.05%.
- The AdaBoost Regressor model yielded an accuracy of around 67.01%.
- Lasso Regression produced an accuracy of approximately 60.24%. - Ridge Regression delivered an accuracy of roughly 69.12%.

Based on experimental analysis, the recommended model is the Ridge Regression model, as it performed the best in terms of optimization.

For future work, we plan to enhance the model's accuracy further by implementing cutting-edge machine learning techniques and evaluating its performance through various methods.

REFERENCES

[1] M. Antonakakis, T. April, M.Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis, D. Kumar, C. Lever, Z. Ma, J. Mason, D. Menscher, C. Seaman, N. Sullivan, K. Thomas, and Y. Zhou, "Understanding the mirai botnet," in Proc. of USENIX Security Symposium, 2017.

[2] Monburinon, Nitis, Prajak Chertchom, Thongchai Kaewkiriya, Suwat Rungpheung, Sabir Buya, and Pitchayakit Boonpou. "Prediction of prices for used car by using regression models." In 2018 5th International Conference on Business and Industrial Research (ICBIR), pp. 115-119. IEEE, 2018.

[3] Hossein Hadian Jazi, Hugo Gonzalez, Natalia Stakhanova, and Ali A. Ghorbani. "Detecting HTTP-based Application Layer DoS attacks on Web Servers in the presence of sampling." Computer Networks, 2017

[4] A. Shiravi, H. Shiravi, M. Tavallaee, A.A. Ghorbani, Toward developing a systematic approach to generate benchmark datasets for intrusion detection, Comput. Security 31 (3) (2012) 357–374.

[5] Z. He, T. Zhang, and R. B. Lee, ―Machine Learning Based DDoS Attack Detection from Source Side in Cloud,‖ in Proceedings of the 2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud), pp. 114–120, New York, NY, USA, June 2017

[6] R. Doshi, N. Apthorpe and N. Feamster, "Machine Learning DDoS Detection for Consumer Internet of Things Devices," 2018 IEEE Security and Privacy Workshops (SPW), San Francisco, CA, 2018, pp. 29-35.

[7] Jerome H. Friedman, (2002), Stochastic gradient boosting, Computational Statistics & Data Analysis, 38, (4), 367-378