LapSense: Leveraging Data Science Forecasts toStrengthen Laptop Shopping

M Abhiram Sharma
Computer Science
KL University
Hyderabad, India

T Venkata Sai Sathvik
Computer Science
KL University
Hyderabad, India

Ankush Kumar
Computer Science
KL University
Hyderabad, India

Vachhani Shivani
Computer Science
KL University
Hyderabad, India

Shaik Abdul Shaan
Computer Science
KL University
Hyderabad, India

Arpita Gupta
Computer Science
KL University
Hyderabad, India

***Abstract*—In a moment when computers are a necessary part of our lives, their costs can differ greatly depending on a broad range of variables. We have created cutting-edge technology for laptop price prediction to aid customers in making wise judgements. To deliver precise and timely pricing estimates for laptops and enable customers to determine the best prices, our project leverages the capability of machine learning and data science. We developed a prognostic model that can analyze this data and provide accurate price forecasts by utilizing cutting-edge machine learning methods. We used Flask, for creating client interfaces, to fabricate a dynamic and user-friendly website that makes this tool available to users.**

***Index Terms*—Laptop, RAM, CPU, GPU, Specifications, Machine Learning, Data Science, Flask, JavaScript.**

I. Introduction

In the realm of laptop purchasing, customers continually grapple with the challenge of obtaining optimal value for their investments [1]. This challenge forms the core of our innovative endeavor – the ”Laptop Price Prediction” project [2][8][9][10]. The modern laptop market offers an overwhelming array of options, leaving buyers facing a pivotal dilemma: how to ascertain genuine value amidst the plethora of choices. The intricacies of evaluating and selecting a laptop that strikes the perfect balance between performance and cost-efficiency have grown increasingly complex [8]. Even when confronted with laptops featuring similar specifications, the pricing of these models can vary drastically, often defying conventional logic [4]. This intricate interplay between feature parity and price diversity perplexes buyers, concealing the underlying factors contributing to these financial discrepancies [5][7]. In essence, the ”Enhancing Laptop Price Prediction through Data Science” project responds to the urgent need for a contemporary, intellectual system that illuminates the path to value-centric laptop acquisitions moreover empower businesses to achieve greater pricing precision. With so many options available to consumers, it might be difficult to choose the best bargain among laptops with comparable specs in the present laptop market. The seemingly equivalent models' varying price spectrum defies common sense, making it difficult to sway what factors are accountable for these cost differences [5]. Because of this a cutting-edge solution that uses data science methods to decipher the complexities of laptop pricing so that businesses and consumers alike can enhance their pricing strategies and make well-informed judgments [7]. The project's objective is to create an intuitive application that estimates laptop prices based on input laptop characteristics. Information like CPU, RAM, storage, GPU, brand, and more will be provided by users [2][8][9][10]. Based on these requirements, we applied ML models to anticipate prices with accuracy [4]. The aim is to flourish a simple and effective application that makes the process of purchasing a laptop easier for customers by enabling them to estimate laptop prices quickly and simply [8]. With this project, people looking for laptops can acquire a reasonable price estimate based on their preferences and a useful and useful solution [9].

II. LITERATURE REVIEW

[1] Predictive Model has become one of the most demanded areas of Information Technology and It has been used effectively in solving many issues in machine learning, for example, Detecting Spam, Medical diagnosis, financial analysis, etc. [2] This paper presents a laptop price prediction system by using the supervised machine learning technique. The research used multiple-linear-regression as the machine learning prediction method which offered 81[3] Machine Learning is seeing its growth more rapidly in this decade. Many applications and algorithms evolve in Machine Learning Day to day. One such application found in journals is house price prediction. [4] One possible computational structure for the upcoming nanocomputing systems is the quantum-dot cellular automata (QCA). It can create circuits with a high space density and low heat dissipation capacity, which makes it possible to create faster computers with less power usage. [5] Data mining is the pro-

cess of removing relevant information from enormous amounts of noisy, fuzzy, incomplete, and random data. Among the most widely used data mining techniques is the decision tree classification method. [6] In the present study, the Artificial Neural Network (ANN), optimized with a Genetic Algorithm (GA-ANN), was employed for seasonal groundwater table depth (GWTD) prediction in the area between the Ganga and Hindon rivers located in Uttar Pradesh State, India. [7] SVM algorithms are becoming more and more popular in scientific and engineering research and applications. [8] The suggested method collects data from a real-time environment and predicts the model's pricing with high accuracy. This study employs Support Vector Regression, Decision Tree Regression and Multi-Linear Regression to forecast laptop prices. [9] The prediction model is delivered with one-of-a-kind combos of features and several regarded computing device learning models. [10] Several classifiers, such as S, Multinomial Logistic Regression, Decision Trees, and Artificial Neural Network are used for the classification task. The classifiers are evaluated with regard to accuracy, recall, precision, and F1-score metrics.

III. METHODS

A. Data Preprocessing

An essential step in preparing data for analysis is data preprocessing. To guarantee that raw data is of high quality and appropriate for statistical analysis or machine learning, it must be cleaned, transformed, and arranged [7]. Managing missing values, scaling features, encoding categorical variables, and eliminating outliers are all part of this process. The accuracy and dependability of the outcomes from any data analysis or modeling task are improved by proper data preprocessing.[10]

1) sampling, the process of choosing a representative subset of data from a large population
2) transformation, which creates a single input by manipulating raw data.
3) denoising, which eliminates data noise.
4) imputation, which fills in mislaid values with statistically significant information.
5) normalization, which arranges information to facilitate quicker access.
6) feature extraction is the procedure for determining a pertinent feature subset that is important in each situation.

TABLE I
COLUMNS ALONG WITH THE DATATYPE OF EACH INDIVIDUAL

Column	Data-type
Company	object
TypeName	object
Inches	float64
ScreenResolution	object
Cpu	object
Ram	object
Memory	object
Gpu	object
OpSys	object
Weight	object
Price	float64

As discussed in Table-I, we observed 9 categorical fields and after checking, we can infer that no null values are also found.
Index(['Company', 'TypeName', 'ScreenResolution', 'Cpu', 'Ram', 'Memory', 'Gpu', 'OpSys', 'Weight'], dtype='object')
Index(['Inches', 'Price'], dtype='object')

We need to split the data of categorical and numerical data into two variables for processing like above. We need to change/transform the categorical data to numerical data for model building.
We have to filter out the well defined and proper unique qualified values, fields like memory, screen resolution are the information that needs to be clubbed into the single entity by using the PPI formula. We also have weight and RAM that are also categorical to be transformed into numerical value that makes it effortless to the formula, we converted the weight replacement as empty string rather than kg, empty string for RAM rather than GB, convert RAM into int32 format and weight into float64 type.

B. Exploratory Data Analysis

Exploratory Data Analysis or popularly called as "EDA" is among the data sciences processes that helps in cleaning,interpreting, and transforming the data in the needed format and concept. EDA helps the data analysts to know about the data they have been working on and how to make it better. [7] Understanding the ambience of the data is very useful now a days to perform visualization or analysis [6].

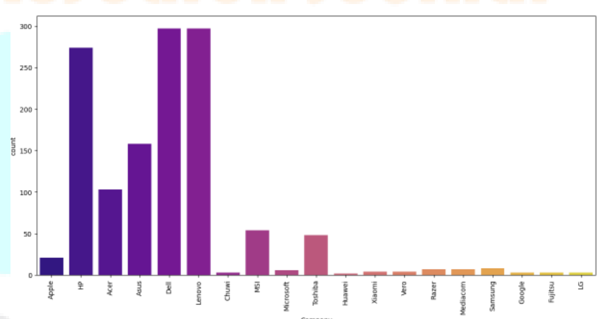

Fig. 1. Bar graph for the count of number of different company laptops

As shown in figure-1, we can see that there are many different companies playing a key role in distribution of laptops for the users where Lenovo, Dell, HP plays a key role at the top.

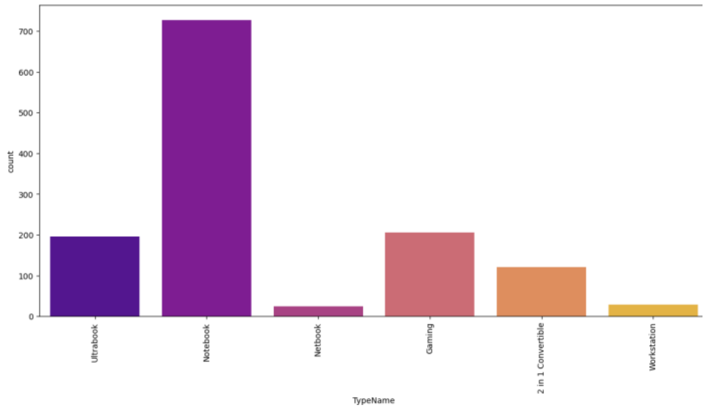

Fig. 2. Bar graph for the count of different type names of the laptops

As shown in figure-2, the different types of laptop that are sold the most are notebook, ultrabook and the Gaming laptop.

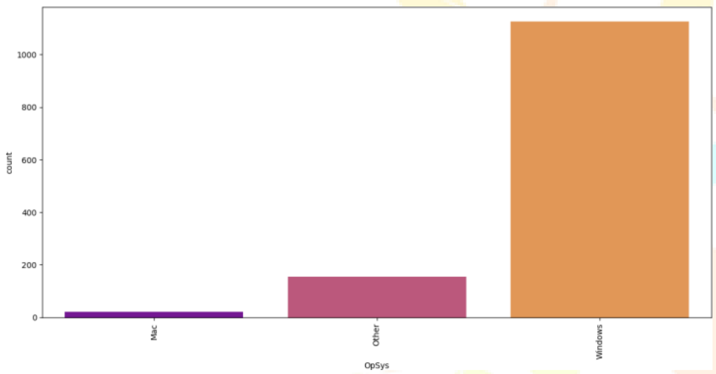

Fig. 3. Bar graph for the count of different type operating systems

As shown in figure-3, Windows operating system is the most famous operating system that is being used world wide among laptops.

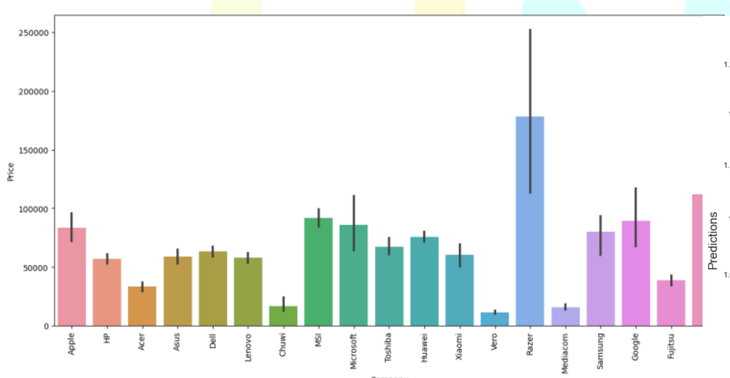

Fig. 4. Bar plot with ticks for the average price of the laptop for each brand

As shown in figure-4, we can observe that the average price of the each laptop brand vs company name that sells laptop.

$$\text{PPI} = \frac{\text{Total Pixels}}{\text{Screen size (inch)}}$$

Fig. 5. PPI formula for the screen resolution generalization

As shown in the figure-5, this is the most important formula as we have screen resolution in width and length as in form 2560X1260 for example, so we need to generalise the term into one variable so that we can use it to calculate the pixels per inch as a variable for easy prediction.

C. Model Building

Our primary step is we had split our data into the training and testing of about 25% testing and 75% training using sklearn test_train_split for validating the model with the right search queries.

Hence, we split the training data as follows:(976, 11) this explains there are gonna be 976 rows of data with all the columns included. We split the testing data as follows: (326, 11) this explains there are gonna be 326 rows of data with all the columns included.

We used our first ML model that is Linear regression which is present inside the sklearn module. Linear regression works on the training data and gives the predicted results according the slope of the equation it is being formed for actual values. After we fit the model and predict the values, i see that incline curved between testing values and predicted values are accurate, the r2_score is 79%.

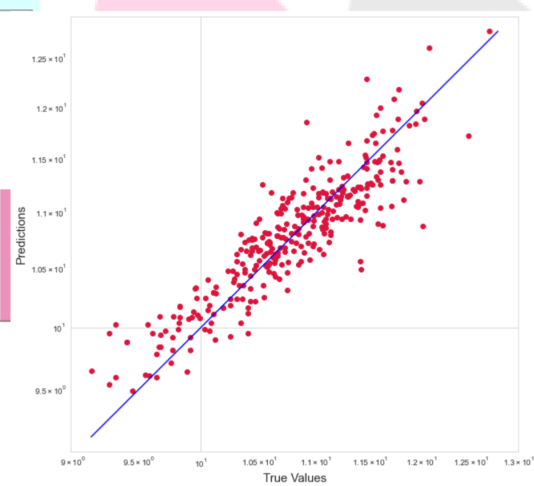

Fig. 6. Scatter plot of true values vs predicted values after linear regression

As shown in the figure-6, we can see that incline curved between testing values and predicted values are accurate.

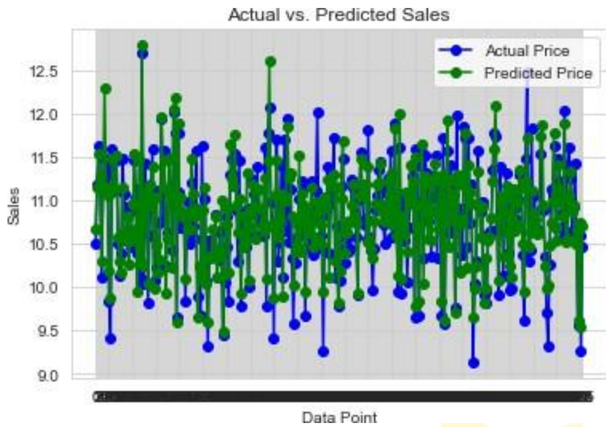

Fig. 7. Bar plot of true values vs predicted values after linear regression

As shown in the figure-7, we can see that bar plot is being plotted between actual values vs predicted values that are almost right on the each data point given they are pretty accurate.

We have then used a ML Model called Lasso regression which is present inside the sk_learn module. The distinction between linear regression and lasso regression is that linear regression aims to fit a linear relationship without considering feature selection, while Lasso regression combines linear modelling with feature selection by penalizing and shrinking the coefficients of less important features to zero, making it a valuable tool for handling high-dimensional datasets and improving model interpretability. After we fit the model and predict the values, i see that incline curved between testing values and predicted values are accurate, the r2_score is 80%.Slightly better than the linear regression.

Fig. 8. Line plot for predicted vs true values after Lasso fit

As shown in the figure-8, we can see that predicted and actual values of the model up here which is plotted on line plot using lasso regression.

Then to make our model more interpretable and accurate we have used the non-linear model called as Random forest regressor which actually performed better than what we have expected.To perform the random forest regressor we need to define some of the hyperparameters called as n_estimators, random_state, max_samples, max_features, max_depth.

To find them we performed the hyper-parameter tuning using GridCV search. Grid Cross-Validation (GridCV), often referred to as Grid-Search is a well-known method used for optimizing ML model with hyper-parameters of a model. Configuration settings that may not be learned are called hyperparameters. from the data but are set prior to training. Examples include the depth of a decision tree, or the regularization strength in a support vector machine.

Random Forest Regressor is a very effective algorithm known for ability to handle intricate data and provide feature importance. It is used for regression tasks in domains like finances, environmental science, and medical fields. When using Random Forest Regressor hyperparameter tuning is a very crucial step to optimize its performance for the application.

After we fit the model and predict the values, I see that the incline curved between testing values and predicted values are accurate, the r2_score is 85%.

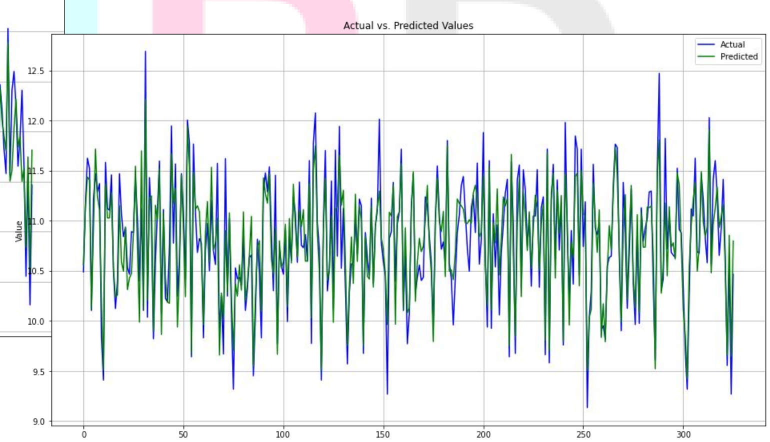

Fig. 9. Line plot for predicted vs true values after Random fit

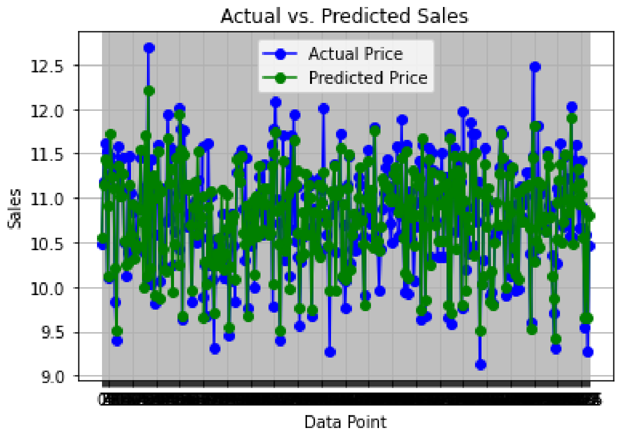

Fig. 10. Bar plot for predicted vs Actual values after Random fit

As shown in the figure-9 and figure-10, we can see that predicted and actual values of the model here plotted looks better than previous algorithms.

Then we have incorporated our model by using pickle and loaded the pickle files into the website which we have developed using flask.

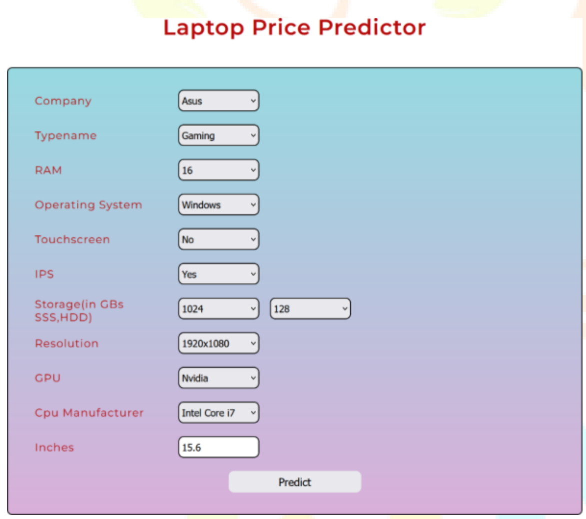

Fig. 11. Website GUI for Laptop price prediction User enhancement

Fig. 12. Predicted price result

As shown in figure-11 and figure-12, We see the website GUI which is designed using flask for User enhancement and the output in figure-12.

IV. Results and Discussions

We achieved a few noteworthy outcomes in the process of assessing our predictive models' performance. At an astounding 79%, the R-squared (R2) score—a gauge of how well our models fit the data—shows that the model accounts for most of the variance in our target variable. Moreover, Lasso regression demonstrated its resilience in managing feature selection and regularization by outperforming even R2 with an accuracy score of 80%. Our Random Forest Regressor model went one step further and attained an accuracy of 85%, showcasing the model's capacity to identify intricate relationships in the data. With an exceptional accuracy of 86%, the Gradient Boosting model outperformed the others and showed great promise for our predictive tasks. In the evaluation of our laptop price prediction model, we achieved an impressive R-squared (R2) score of 86% using the Random Forest Regressor. This score signifies that our model explains approximately 86% of the variance in the laptop prices, indicating a strong ability to capture the relationships between the features we considered and the actual prices.Our model is quite reliable in making price predictions for laptops. An R2 score of 86% demonstrates the effectiveness of our Random Forest Regressor in accurately estimating laptop prices and makes it a promising tool for assisting consumers in making well-informed purchasing decisions in the dynamic laptop market.

V. Conclusion

We began by researching existing systems for price prediction models. Next, we gathered a variety of sales datasets and crafted our own dataset, tailoring it to our specific needs with custom columns and rows. After that, we dove into pre processing and exploratory data analysis, aiming to extract valuable features and identify any data outliers. We effectively removed these outliers and used Pandas and lambda functions to extract essential data from relevant columns. We split our processed data into a 70-30 train-test ratio. To create accurate predictions, we applied Linear Regression, Lasso Regression, and Random Forest Regression algorithms. We carefully compared the model accuracies and ultimately selected the Random Forest model as our best performer. Lastly, we developed a user-friendly webpage using Flask to render front-end, enabling users to input data and receive price predictions.

References

[1] A. A. Gupta and S. Vijaykumar, "Mobile price prediction by its features using predictive model of machine learning," Studies in Indian Place Names, vol. 40, no. 35, pp. 906-913, 2020.

[2] V. Surjuse, S. Lohakare, A. Barapatre, and A. Chapke, "Laptop Price Prediction using Machine Learning," 2022.

[3] M. Thamarai and S. P. Malarvizhi, "House Price Prediction Modeling Using Machine Learning," International Journal of Information Engineering & Electronic Business, vol. 12, no. 2, 2020.

[4] S. Afroz and N. J. Navimipour, "Memory designing using quantum dot cellular automata: systematic literature review, classification, and current trends," J. Circuits Syst. Comput., vol. 26, no. 12, pp. 1730004, 2017.

[5] H. Sharma and S. Kumar, ”A Survey on Decision Tree Algorithms of Classification in Data Mining,” International Journal of Science and Research, vol. 5, no. 4, pp. 2094–2097, 2016, doi: 10.21275/v5i4.nov162954.

[6] N. Kumar and D. Kumar, ”Classification using Artificial Neural Network Optimized with Bat Algorithm,” International Journal of Innovative Technology and Exploring Engineering, vol. 9, no. 3, pp. 696–700, 2020, doi: 10.35940/ijitee.c8378.019320.

[7] J. Cervantes, F. Garcia-Lamont, L. Rodriguez-Mazahua, and A. Lopez, ”A comprehensive survey on support vector machine classification: Applications, challenges, and trends,” Neurocomputing, vol. 408, pp. 189-215, 2020, DOI: 10.1016/j.neucom.2019.10.11824.

[8] C. L. Reddy, K. B. Reddy, G. R. Anil, S. N. Mohanty, and A. Basit, ”Laptop Price Prediction Using Real Time Data,” 2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC), pp. 1-5, 2023.

[9] M. A. Shaik, M. Varshith, S. S. Vyshnavi, N. M. S. Jana, and R. Sujith, ”Laptop Price Prediction using Machine Learning Algorithms,” 2022 International Conference on Emerging Trends in Engineering and Medical Sciences (ICETEMS), pp. 226-231, 2022.

[10] A. A. Syed, Y. H. Lukas, and A. Wibowo, ”A Comparison of Machine Learning Classifiers on Laptop Products Classification Task,” Lecture Notes in Engineering and Computer Science: Proceedings of The International Multiconference of Engineers and Computer Scientists, pp. 104-110, 2021.