## National College of Ireland

## Project Submission Sheet – 2023

**Student Name:** Aravind Hallimysore Kalegowda, Rohan Kanagal Sathyanarayana

**Student ID:** x22104275, x19203829

**Programme:** MSc Data Analytics **Year: 2023**

**Module:** Domain Application of Predictive Analytics (MSCDAD_JAN23A_I)

**Lecturer:** Qurrat Ul Ain

**Submission Due Date: 11/08/2023**

**Project Title:** Exploring and analysing, market trend and specifications to predict laptop price

**Word Count:** 2604

**I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.**
**ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.**
**Signature:**

**Date: 10/08/2023**

**PLEASE READ THE FOLLOWING INSTRUCTIONS:**

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

**Office Use Only**
Signature:
Date:
Penalty Applied (if applicable):

# Exploring and analyzing, market trend and specifications to predict laptop price.

1st Aravind Hallimysore Kalegowda
*MSc in Data Analytics*
*National College of Ireland*
Dublin
x22104275@student.ncirl.ie

2nd Rohan Kanagal Sathyanarayana
*MSc in Data Analytics*
*National College of Ireland*
Dublin
x19203829@student.ncirl.ie

*Abstract*—This project aims to develop a predictive model using machine learning techniques to forecast laptop prices based on various specifications. The analysis involves data preprocessing, feature engineering, and utilization of the Random Forest Regressor model to provide valuable insights into the factors influencing laptop prices. The results of this study have implications for consumers, manufacturers, and retailers in the competitive laptop market, aiding them in making informed decisions and optimizing pricing strategies.

*Index Terms*—Laptop Price Prediction, Machine Learning, Random Forest Regressor, Predictive Analysis, Laptop Specifications

## I. INTRODUCTION

This project involves using domain knowledge and predictive analysis by using a machine learning model that can predict laptop prices based on different specifications. With laptops being an essential part of our daily lives this project holds significant importance for consumers, manufacturers and retailers in the competitive laptop market. By leveraging a comprehensive dataset containing information on different laptop models and their corresponding prices, the project involves data preprocessing, feature engineering, and data visualization to gain insights into the factors that influence laptop prices, such as brand, screen size, processor, memory, and more. Utilizing the Random Forest Regressor model the predictive analysis will enable accurate price forecasting, providing valuable information for informed decision making and pricing strategies. The ultimate goal of this project is to offer practical implications and actionable insights for industry stakeholders, enabling them to optimize pricing dynamics and meet customer demands effectively.

## II. LITERATURE REVIEW

In this paper [1] by Chada Lakshma Reddy, K Bhargav Reddy and their colleague the research illustrates that due to the global pandemic sales of laptop through online have risen to the highest level ever. Work from home as well as online education and various other tasks requires the utilisation of a laptop. A feature based approach method benifits the consumer for taking a good decision to purchase the laptop. In this assessment the researchers has presented an approach for estimating the price of laptop using real time data collected from a website that sells goods online. For prediction of laptop prices they have utilized Support Vector Regression, Decision Tree Regression, and Multi-Linear Regression. Wherein , the linear regression method acquired 59 percent accuracy , support vector regression technique achieved an overall accuracy of 87.5percent , the reason for this is because due to the fact that most regression coefficients tries to fit a straight line to the data by minimizing the cost function but the SVM eventually end up by fitting a curve instead of a straight line , In conclusion the researchers suggested that Decision Tree Regression outperformed between these two models with an accuracy of 93percent and while suggesting their view on future work they mentioned that if there is a data which is sufficient to build a deep learnig model a technique that predicts the best company along with the price can be built.

In this paper [2] by Astri Dahlia Siburian, Daniel Ryan Hamonangan Sitompul and their colleagues they stated that following the covid-19 pandemic numerous tasks and activities have been carried out from home(WFH). In accordance with the information gathered by East Java Central Statistics Agency (BPS), major and medium sized companies which opted to work from home are 32.37 percent in 2021, with this statistics numerous individuals need a work device(at this instance , Laptop) to improve their level of productivity , hence the researchers expressed that, one need a laptop with features that is suitable in increasing the efficiency in their productivity. Further they expressed that ,to prevent consumers overspending on laptops, an approach to estimate the cost of laptop according to given or required specifications needs to be developed and developed a model using machine learning algorithms like Random forest regressor , Gradient boosting regressor and XG boost regressor where they achieved an accuracy of 80.79 percent , 90.55 percent and 92.77 percent respectively . Finally they concluded that XGBoost was the model featuring the highest R2 value and lowest RMSE value than Random Forest and
Gradient Boosting.

In this research paper [3] it has shown how people buy things and they found some interesting insights. After studying different factors that can influence their buying choices such as

the environment the company they are buying from their personal preferences and how they interact with others researcher used a powerful computer program called Random Forest to make predictions. The results were impressive with an accuracy rate of 94percent. This means that researcher predictions were quite close to the actual behaviour of customers when it comes to buying products. By using Random Forest researcher was able to give more weightage to the numerical results and gain a better understanding of customer behaviour which can be valuable for businesses to improve their marketing strategies and cater to their customers needs more effectively.

In this paper [4] the researchers have assessed the sale of used(second-hand) electronic devices as well as its significance. Considering the rapid growth of data generation on business of electric goods along with the availability of information to the public the researchers made use of the opportunity to web scrape the historical information about the sales of the electronic goods whereby then utilized the data to present the predictive model. They employed three distinct algorithms of machine learning in order to determine the price of the used electronic products like laptop, cellphones etc. They utilized Random forest , Linear Regression and Multi-Layer Perceptron in their assessment where finally they concluded that RF model performs better than either of the two algorithms when it comes to predicting the errors. In order to ensure the reliability and accuracy the team of researchers carried out the k-fold cross-validation along with hyperparameter tuning with the strategy of grid search.

## III. METHODOLOGY

In this project we follow the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. We started with understanding the business objectives of predicting laptop prices. Next we gathered and explore the laptop data, perform data preprocessing and engineer relevant features. Then we selected the Random forest Regressor algorithm, trained the model and evaluated its performance using metrics like R-squared and MAE. And this implementation can be deployed the model for practical use and ensure its maintenance for continued accuracy.

## IV. RESEARCH AND INVESTIGATION INTO THE APPLICABLE TECHNIQUES

In this project we conducted thorough research and investigation into various applicable techniques for predicting laptop prices. Our primary goal was to identify the most effective and accurate approach that could provide valuable insights for businesses in the competitive laptop market.

During the research phase we explored several machine learning algorithms including Random Forest, XGBoost, and Decision Tree. We analyzed their strengths, weaknesses and suitability for our specific dataset and prediction task. And then we reviewed existing literature and studies related to laptop price prediction and consumer behavior to gain a deeper understanding of the subject.

To ensure the success of our investigation we focused on data preprocessing and exploratory data analysis (EDA). Cleaning and preparing the dataset were crucial steps to eliminate any inconsistencies and ensure the quality of our analysis. EDA helped us uncover essential patterns and trends in the data revealing significant features that influenced laptop prices.

Based on our research findings we proceeded with the implementation of the selected techniques. We built predictive models using Random Forest, XGBoost and Decision Tree algorithms. Each model was carefully fine tuned by optimizing hyperparameters to achieve optimal performance. We then evaluated these models using various metrics to assess their accuracy in predicting laptop prices.

We were able to determine the best appropriate technique through extensive research and inquiry that would not only produce reliable predictions but also significant business insights for organisations in the laptop sector. This information can help consumers, sellers and manufacturers to make informed decisions to buy laptop and expand their product sales and market competitiveness.

## V. IMPLEMENTATION OF THE SELECTED TECHNIQUE

We used the Random Forest Regressor (RFR) technique for this project since it can effectively handle both numerical and categorical input. Our dataset contained a mix of numerical features like RAM, weight, and price as well as categorical attributes like brand, operating system, and screen resolution. RFR's capacity to handle diverse data types without extensive preprocessing made it a suitable choice saving us time and effort. Moreover RFR's ensemble approach reduced the risk of overfitting given the large number of features and potential interactions. This ensured better generalization to unseen data and improved the models reliability. And also RFR is built in feature importance analysis allowed us to identify key determinants affecting laptop prices providing valuable insights for businesses seeking to optimize their product offerings and marketing strategies. Apart from RFR we also utilized essential techniques such as data preprocessing and exploratory data analysis (EDA) to prepare and understand the dataset.

*A. Data preprocessing*

Data preprocessing involved checking missing values, understanding the what variables we have and how dataset looks by using head() function and understanding type of variable by using info() functionand encoding categorical variables as required. We can find the Null value plot in Fig.1 and here we can see that in dataset there is no null values.

In Fig.2 it has given the summary of the dataset by using the describe() function. It has given the key statistical information about the numerical columns in the DataFrame. This will help us to understand the central tendencies, spreads, and overall distribution of your data.

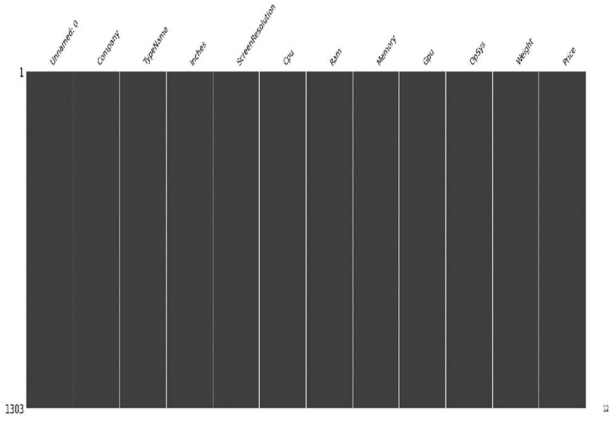


Fig. 1. Null Value plot

```
In [12]: #Summary Statistics of our Dataset
         df_laptop.describe()
```

Out[12]:

| | Unnamed: 0 | Inches | Price |
|---|---|---|---|
| count | 1303.00000 | 1303.000000 | 1303.000000 |
| mean | 651.00000 | 15.017191 | 59870.042910 |
| std | 376.28801 | 1.426304 | 37243.201786 |
| min | 0.00000 | 10.100000 | 9270.720000 |
| 25% | 325.50000 | 14.000000 | 31914.720000 |
| 50% | 651.00000 | 15.600000 | 52054.560000 |
| 75% | 976.50000 | 15.600000 | 79274.246400 |
| max | 1302.00000 | 18.400000 | 324954.720000 |

Fig. 2. Summary of the data

In Fig.3. It give the information about several essential transformations on the laptop dataset. We removed the 'kg' unit from the 'Weight' column, converting it into numerical values for uniformity. And also we transformed the median prices for CPUs and GPUs using the 'Price' column, creating 'CPU-avg' and 'GPU-avg' features. These features gives insights into average pricing trends for CPUs and GPUs. Further enhancing our dataset, we categorized CPUs and GPUs into groups based on their average prices adding cpu-group and gpu-group features. These changes found the base for more significant analysis and modelling in our project later stages.

```
In [16]: #Remove 'kg' from the 'Weight' column in df_laptop  dataframe.
         df_laptop['Weight'].replace(to_replace='kg', value='', regex=True, inplace=True)

In [17]: #Convert 'Weight' values to float data type in df_laptop dataframe
         df_laptop['Weight'] = df_laptop['Weight'].astype("float64")

         Grouping Variables

In [18]: #Calculate the median price for each unique 'Cpu' and 'Gpu' group and assign the values to 'CPU_avg' and 'GPU_avg' columns in df_
         df_laptop['CPU_avg'] = df_laptop.groupby(['Cpu'])['Price'].transform('median').round(2)
         df_laptop['GPU_avg'] = df_laptop.groupby(['Gpu'])['Price'].transform('median').round(2)

In [19]: #Applying 'group' function to 'CPU_avg' and 'GPU_avg' columns, and creating 'cpu_group' and 'gpu_group' columns based on the resu
         df_laptop['cpu_group'] = df_laptop.apply(lambda x: group(x['CPU_avg']),axis=1)
         df_laptop['gpu_group'] = df_laptop.apply(lambda x: group(x['GPU_avg']),axis=1)
```

Fig. 3. Transforming variables

*B. Exploratory Data Analysis*

After the data preprocessing next step of this project was EDA. EDA helped us to gain a deeper understanding of the dataset with help of identifying patterns, spot differences and make informed assumptions. In this project various plots were created to understand the distribution of Price and its relationship with other features. In this project bar plots were used to visualize the relationship between 'Price' and 'Company', 'TypeName', 'Ram', 'OpSys', 'Touchscreen', and 'Inches'. These steps were crucial in ensuring the dataset quality and preparing it for effective model training.

We were able to create a stable and accurate predictive model for laptop price prediction by using RFR and supplementing it with proper data preprocessing and EDA. Our research was effective in handling the complexities of the dataset and producing valuable insights for firms in the laptop industry due to the combination of RFRs adaptability, robustness and feature importance analysis as well as other preliminary methodologies. This projects overall methodology facilitated a full understanding of the factors driving laptop prices and their impact on consumer behaviour.

*C. Data Transformation Techniques*

Data transformation is another vital step to prepare the data for modeling. In Fig.4. The target variable 'Price' was log-

transformed to achieve a more normal distribution, which is a common practice for dealing with skewed continuous variables. Moreover categorical features were label-encoded to convert them into numerical form. This was necessary because machine learning algorithms typically require numerical input.

Log-Normal Transformation

```
In [33]: #Distribution plot of the logarithm of 'Price' column.
         sns.distplot(np.log(df_laptop['Price']))
         plt.show()
```

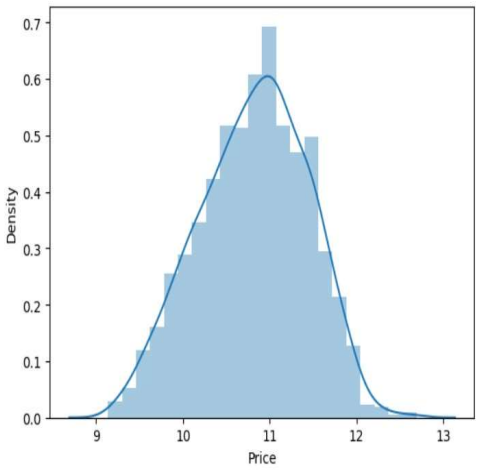


Fig. 4. Log-Normal Transformation

## VI. The findings, in terms of both the quantitative results and their business value qualitative interpretation

### *A. Quantitative Results*

The quantitative evaluation of our project's Random Forest Regressor (RFR) model demonstrates its predictive ability. In Fig.5. we see that the model received high training and test scores of 86.26 percent and 80.77 percent, respectively. This demonstrates the model's capacity to detect underlying patterns in the data, resulting in correct predictions. The Mean Absolute Error (MAE) of about 0.21 and the Mean Squared Error (MSE) of about 0.07 demonstrate the model's efficacy in minimising forecast deviations. Furthermore, the Root Mean Squared Error (RMSE) of around 0.27 represents the average magnitude of prediction mistakes, offering vital information into the model's precision. The R-squared (R2) value of 0.81 indicates the models capacity to explain the variance in the target variable, indicating its dependability in projecting laptop pricing. These quantitative metrics collectively demonstrate the RFR models strong performance and its potential to enhance business decision making.

Running Random Forest

```
In [43]: #RandomForestRegressor model with the specified hyperparameters
         from sklearn.ensemble import RandomForestRegressor
         regressor_rf = RandomForestRegressor(n_estimators = 100, min_samples_split = 2,
         max_depth= 7,  criterion = 'squared_error', random_state = 0)
         regressor_rf.fit(X_train, y_train)
         normal_rf = regressor_rf.score(X_train, y_train)
         normal_rf_test = regressor_rf.score(X_test, y_test)
         previsoes = regressor_rf.predict(X_test)
         mae_normal_rf  = mean_absolute_error(y_test, previsoes)
         mse_normal_rf = mean_squared_error(y_test, previsoes)
         rmse_normal_rf = np.sqrt(mean_squared_error(y_test, previsoes))
         r2_normal_rf = r2_score(y_test, previsoes)

         print('Train : ', normal_rf)
         print('Test :', normal_rf_test)
         print('MAE :', mae_normal_rf)
         print('MSE :', mse_normal_rf)
         print('RMSE :', rmse_normal_rf)
         print('R2 :', r2_normal_rf)

         Train :  0.8626290068555048
         Test : 0.8077461262928047
         MAE : 0.213657384893425 95
         MSE : 0.07338491346016798
         RMSE : 0.27089649953472633
         R2 : 0.8077461262928047
```

Fig. 5. Random forest modeling

### *B. Feature Importances*

Using the Random Forest Regressor we depicted feature importances through a bar chart. This visual representation in Fig.6. showcased the relative weight of various laptop attributes in determining their prices. The resulting bar chart clearly illustrates the relative significance of each feature, assisting businesses in pinpointing key drivers of pricing decisions and optimizing their strategies accordingly.

*1) Predictive Results using Random Forest Regressor:* In Fig.7. we present the predictive results generated by our implemented Random Forest Regressor (RFR) model. This model was precisely trained and fine tuned to estimate laptop prices accurately. The following table showcases a subset of our predictions alongside the actual and predicted prices for comparison. It is important to note that the 'real price'

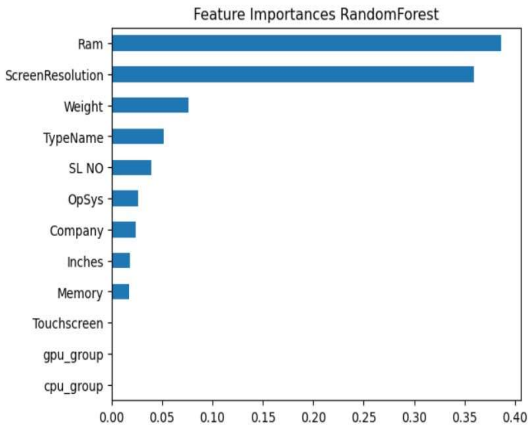


Fig. 6. Random forset modeling

represents the actual laptop price, while 'rf pred' corresponds to the predicted price provided by our RFR model.

| Index | SL NO | real price | rf pred |
|---|---|---|---|
| 0 | 248 | 35964.0000 | 26138.976921 |
| 1 | 556 | 13586.4000 | 21705.122407 |
| 2 | 693 | 137995.2000 | 78683.570123 |
| 3 | 387 | 72940.3200 | 97089.385076 |
| 4 | 781 | 125208.0000 | 111819.195954 |
| 5 | 379 | 45074.8800 | 47529.260307 |
| 6 | 716 | 62817.1200 | 57813.851394 |
| 7 | 880 | 90576.0000 | 47574.182943 |
| 8 | 654 | 19441.8720 | 20312.250782 |
| 9 | 994 | 47365.9200 | 57813.851394 |
| 10 | 825 | 26586.7200 | 39792.046600 |
| 11 | 743 | 53226.7200 | 57940.519020 |
| 12 | 1243 | 95850.7200 | 90030.903367 |
| 13 | 335 | 67559.0400 | 69487.321706 |
| 14 | 708 | 63456.4800 | 57940.519020 |
| 15 | 235 | 47898.7200 | 86155.131872 |
| 16 | 332 | 71075.5200 | 50558.244603 |
| 17 | 1139 | 115709.2416 | 64269.909519 |
| 18 | 739 | 21887.4240 | 25531.117139 |
| 19 | 573 | 44222.4000 | 57601.057962 |
| 20 | 270 | 104370.1920 | 79666.364007 |

Fig. 7. Predicted Results

*C. Qualitative results*

In our Random Forest Regressor model not only has high predictive potential but also provides significant business value. Businesses may make well informed decisions when determining product prices by effectively projecting laptop prices, matching them with market trends and customer preferences. This price precision can lead to higher customer satisfaction and brand loyalty. The Random Forest Regressor implementation also capitalises on feature importance allowing us to find the most relevant aspects in deciding laptop prices. This knowledge can help organisations improve product features that have the greatest influence on purchasing decisions. And also the models dependability and applicability in real world circumstances are ensured by the incorporation of rigorous data preprocessing and exploratory analysis. In summary our project offers a strong tool that enables firms to optimise their pricing strategies and successfully react to changing consumer needs.

An also our models ability to predict laptop pricing helps consumers, sellers with transparency and trust in their purchasing and selling decisions. It provides customers with a clear understanding of how specific features contribute to the final price, allowing them to make value driven decisions. This transparency creates trust between organisations and consumers which is critical in today's competitive market context.

## CONCLUSION

In conclusion the quantitative excellence of our RFR model aligns seamlessly with its qualitative business value. The accurate price predictions and the insights into feature importance enhance decision making processes for both businesses and consumers. This collaboration between technological precision and practical utility highlights the significance and application of our study findings in the dynamic arena of laptop pricing and customer behaviour.

This project has the potential to change how businesses decide on prices. Using a machine learning model called Random Forest Regressor (RFR) we help businesses make better choices about how much to charge for their products. This can give them an advantage in the market. Getting the pricing right doesn't just make more money it also keeps customers happy and loyal. And we have used smart methods to handle the data and look at it closely making sure our predictions are really good. In a world where customers habits are always changing this project gives businesses a smart way to use data and do well.

## REFERENCES


[1] Chada, L. R., Reddy, K. B., Anil, G. R., Mohanty, S. N., and Basit, A. (2023). "Laptop Price Prediction Using Real Time Data.", (23-25 January 2023). Link: https://ieeexplore.ieee.org/abstract/document/10085473

[2] Astri Dahlia Siburian, Daniel Ryan Hamonangan Sitompul, Stiven Hamonangan SinuratAndreas Situmorang, Ruben, Dennis Jusuf Ziegel, Evta Indra, " Laptop Price Prediction with Machine Learning Using Regression Algorithm", Link:http://jurnal.unprimdn.ac.id/index.php/JUSIKOM/article/view/2850/1879.

[3] Harsh Valecha; Aparna Varma; Ishita Khare; Aakash Sachdeva; Mukta Goyal. "Prediction of Consumer Behaviour using Random Forest Algorithm", Link: https://ieeexplore.ieee.org/abstract/document/8597070

[4] Muhammad Hasnain , Abdul Sajid , M. Arshad Awan. "Predicting The Price Of Used Electronic Devices Using Machine Learning Techniques" ,'2023-07-18' Link: http://ijcrt.smiu.edu.pk/ijcrt/index.php/smiu/article/view/152

## Individual Contribution of this Project:

**Aravind :** EDA ,Visualizations , Data Transformation , Modelling , Result Analysis and report writing were divided and well-coordinated between both of us .

**Rohan :** Dataset choosing , Data Exploration , Data Cleaning , Feature engineering along with this few models building and report writing were divided and well-coordinated between both.