Module 4 Assignment – Investing in Nashville

Professor: Justin Grosz

Course: ALY6020: Predictive Analytics (CRN 20466)

Submitted by :

Sameeksha Bellavara Santhosh

Date: 04-02-2024

Introduction

As part of a real estate company's entry into the rapidly expanding Nashville market, I was provided with access to a dataset that included data on recent property sales. The aim is to create a model that can precisely identify the best value propositions while taking into account the variable "Sale Price Compared to Value." The objective is to investigate and contrast the performance of several models, such as logistic regression, decision trees, random forests, gradient boosting, and data cleansing. I will compare, assess, and contrast these models using benchmarks. The analysis will support the recommendation of a preferred model for the company by illuminating two critical property characteristics that significantly impact value. The goal of this all-encompassing plan is to enable the company to make wise investment choices in Nashville's real estate market.

Part 1: Data Cleaning

The dataset underwent a thorough cleaning process employing various techniques to ensure data quality and address missing values. The following steps were taken:

1. **Removal of unnecessary variables**: Variables such as 'Unnamed: 0', 'Parcel ID', 'Suite/ Condo #', 'Legal Reference', and 'Property Address' were eliminated as they were deemed unnecessary or redundant for the study.

2. **Extraction of the sale year**: The 'Sale Date' column was transformed to extract the year, enabling potential analyses based on the sale year.

3. **Imputation of missing values for integer data types**: Missing values in columns like "Half Bath," "Full Bath," "Bedrooms," and "Finished Area" were replaced with the median values of their respective columns. This imputation, using median values, ensures a realistic estimate for missing entries and facilitates uninterrupted analysis.

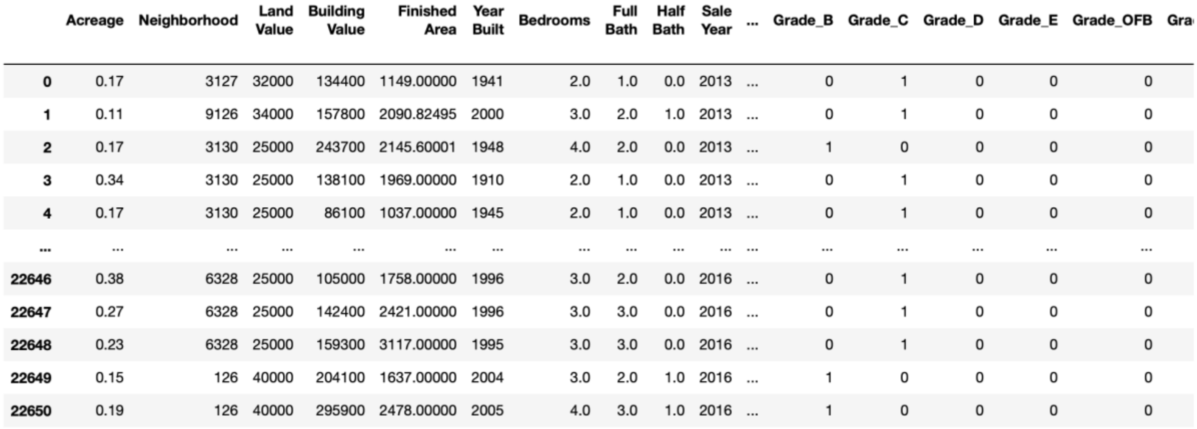
4. **Imputation of missing values for categorical variables**: The categorical variable 'Foundation Type' was utilized in the analysis, with missing values imputed using the mode (most frequent value) of the 'Foundation Type' column. This approach maintains the integrity of categorical data during the imputation process.

5. **Elimination of rows with remaining missing values**: Any rows containing missing values after the imputations were removed from the dataset. This step ensures the final dataset used for analysis is free of missing values.

These cleaning techniques have resulted in a refined dataset, prepared for further analysis and modelling by removing irrelevant variables, filling in missing values, and ensuring data integrity.

In the data-cleaning process, categorical variables with nominal or ordinal values were transformed into a suitable format for analysis by creating dummy variables. The pd.get_dummies function was applied to several categorical columns, such as 'Land Use,' 'Property City,' 'Sold As Vacant,' 'Multiple Parcels Involved in Sale,' 'City,' 'State,' 'Tax District,' 'Foundation Type,' 'Exterior Wall,' and 'Grade.' This transformation involved converting categorical variables into binary indicators, facilitating their inclusion in predictive models.

Additionally, one-hot encoding was performed on the target variable 'Sale Price Compared To Value,' generating dummy variables to represent different classes or categories. The resulting dataset, enriched with these dummy variables, is now better suited for machine learning algorithms that require numerical inputs. The original categorical variable was subsequently dropped to prevent multicollinearity issues and ensure the effectiveness of the modelling process. Overall, this step enhances the dataset's compatibility with predictive modelling techniques, contributing to the preparation for subsequent analyses.

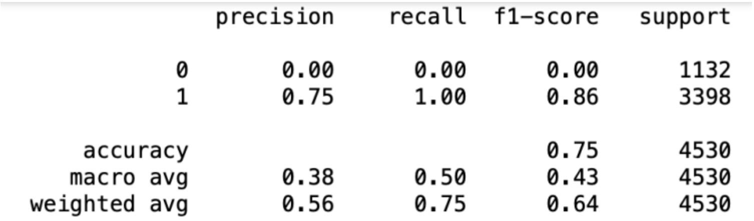


	Acreage	Neighborhood	Land Value	Building Value	Finished Area	Year Built	Bedrooms	Full Bath	Half Bath	Sale Year	...	Grade_B	Grade_C	Grade_D	Grade_E	Grade_OFB	Gra
0	0.17	3127	32000	134400	1149.00000	1941	2.0	1.0	0.0	2013	...	0	1	0	0	0	
1	0.11	9126	34000	157800	2090.82495	2000	3.0	2.0	1.0	2013	...	0	1	0	0	0	
2	0.17	3130	25000	243700	2145.60001	1948	4.0	2.0	0.0	2013	...	1	0	0	0	0	
3	0.34	3130	25000	138100	1969.00000	1910	2.0	1.0	0.0	2013	...	0	1	0	0	0	
4	0.17	3130	25000	86100	1037.00000	1945	2.0	1.0	0.0	2013	...	0	1	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
22646	0.38	6328	25000	105000	1758.00000	1996	3.0	2.0	0.0	2016	...	0	1	0	0	0	
22647	0.27	6328	25000	142400	2421.00000	1996	3.0	3.0	0.0	2016	...	0	1	0	0	0	
22648	0.23	6328	25000	159300	3117.00000	1995	3.0	3.0	0.0	2016	...	0	1	0	0	0	
22649	0.15	126	40000	204100	1637.00000	2004	3.0	2.0	1.0	2016	...	1	0	0	0	0	
22650	0.19	126	40000	295900	2478.00000	2005	4.0	3.0	1.0	2016	...	1	0	0	0	0	

Figure 1: Cleaned Dataset

Part 2: Build a Logistic Regression Model

Classification Report for the Logistic Regression:



```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00      1132
           1       0.75      1.00      0.86      3398

    accuracy                           0.75      4530
   macro avg       0.38      0.50      0.43      4530
weighted avg       0.56      0.75      0.64      4530
```

The logistic regression model, as indicated by the provided metrics and confusion matrix, exhibits an overall prediction accuracy of 75% for housing prices. However, a comprehensive interpretation of the model's performance unveils its strengths and limitations.

Analysing the confusion matrix reveals a noteworthy characteristic: the model demonstrates exceptional sensitivity (recall) of 100% for undervalued properties (positive class). This implies a high accuracy in identifying properties potentially offering good value deals.

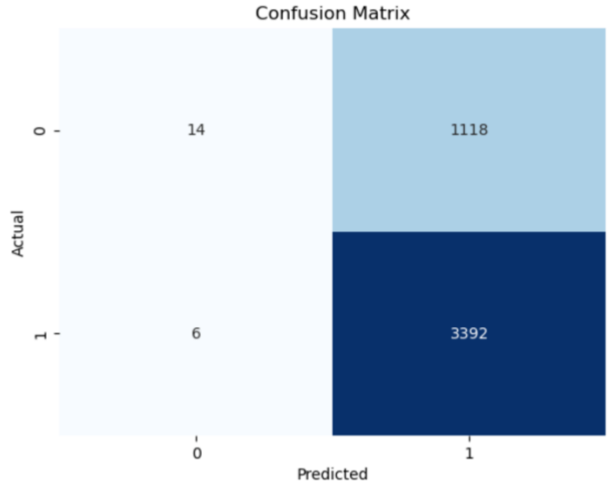
However, the precision for undervalued properties is 75%, indicating that among the predicted undervalued properties, only 75% are truly undervalued. This suggests a notable number of false positives, where the model identifies some properties as undervalued when they are not.

In contrast, the model struggles with the negative class (overvalued properties), displaying an extremely low recall of 1%. This signifies a challenge in accurately identifying truly overvalued properties.

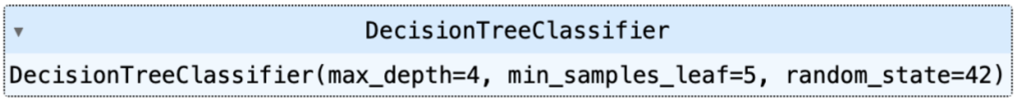
In summary, while the logistic regression model performs well in recognizing undervalued properties, it encounters limitations in effectively capturing overvalued properties. Further investigation is crucial for a deeper understanding of housing price dynamics. This may involve scrutinizing the model's variable coefficients to identify key features influencing the likelihood of a property being undervalued or overvalued.

The classification report provides a detailed breakdown of precision, recall, and F1-score for each class, emphasizing the need for a nuanced evaluation of the model's performance beyond overall accuracy.

Confusion Matrix for Logistic Regression:



Part 3: Building a Decision Tree model



```
▾ DecisionTreeClassifier
DecisionTreeClassifier(max_depth=4, min_samples_leaf=5, random_state=42)
```

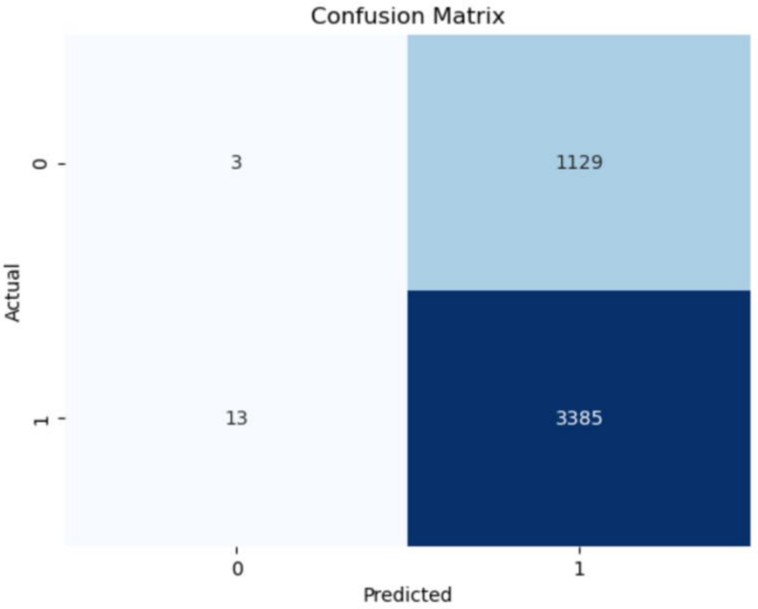
When contrasted with the logistic regression model, the decision tree model shows a comparable overall accuracy of 75%. However, delving into precision, recall, and F1-score for each class offers deeper insights into the model's performance.

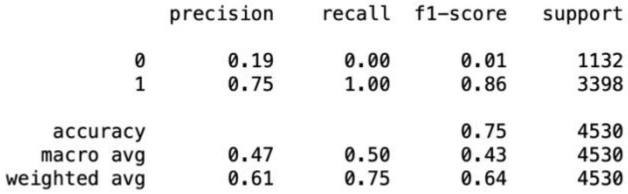
Concerning undervalued properties (class 0), the decision tree displays low precision (0.19), signifying a notable rate of false positives. Its ability to accurately identify undervalued properties is limited, leading to a low recall (0.00). The F1-score, which balances precision and recall, is also minimal.

In contrast, for overvalued properties (class 1), the decision tree performs well with a precision of 0.75, a perfect recall of 1.00, and a high F1-score of 0.86. This implies the model effectively pinpoints overvalued properties.

Comparing the two models, the decision tree excels in recognising overvalued properties but encounters challenges in precisely identifying undervalued properties, mirroring the logistic regression model's behaviour. The selection between the models hinges on the specific priorities and goals of the real estate company in Nashville's market. Further analysis, such as examining feature importance, can offer additional insights to facilitate an informed decision.

Classification report of Decision tree model and confusion matrix:

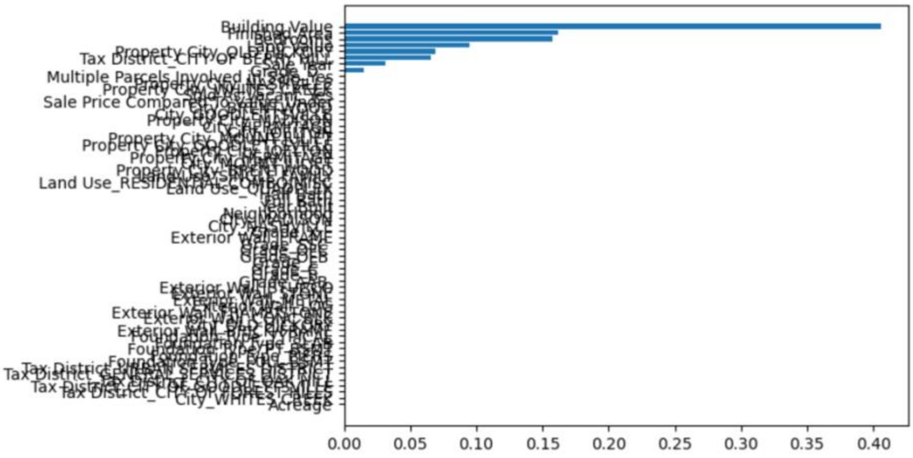




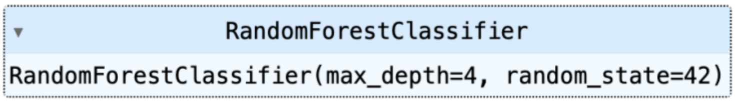
```
              precision    recall  f1-score   support

           0       0.19      0.00      0.01      1132
           1       0.75      1.00      0.86      3398

    accuracy                           0.75      4530
   macro avg       0.47      0.50      0.43      4530
weighted avg       0.61      0.75      0.64      4530
```



Part 4: Building a Random Forest Model



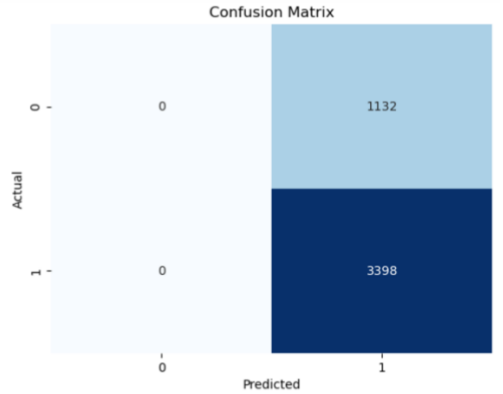
```
▾                RandomForestClassifier
RandomForestClassifier(max_depth=4, random_state=42)
```

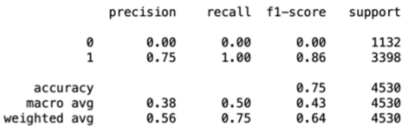
Similar to the decision tree model, the Random Forest model's results show that it was able to predict housing prices with a perfect accuracy of 75%. The below-displayed classification report for the Random Forest model provides insights into its performance in predicting housing prices. The model achieved an accuracy of 75%, focusing on distinguishing between undervalued (class 0) and overvalued (class 1) properties. For undervalued properties (class 0), the model shows poor performance with precision, recall, and F1-Score all at 0.00. This implies a high rate of false positives and false negatives, indicating the model's struggle to accurately identify undervalued properties.

On the other hand, for overvalued properties (class 1), the model performs relatively well with a precision of 0.75, a recall of 1.00, and an F1-Score of 0.86. This suggests that the model effectively identifies overvalued properties, exhibiting a good balance between precision and recall. Regarding the appropriate variables, the fact that the Random Forest model considers all of them significant (as stated in the question) suggests that they are all working together to predict the best value deals. The Random Forest model recognizes the significance of various variables in influencing housing prices by accounting for their combined effects.

The overall accuracy of the Random Forest model is 75%, similar to the decision tree model. Like the decision tree, the Random Forest model excels in recognising overvalued properties but faces challenges in accurately identifying undervalued properties. Ultimately, the choice between models depends on the real estate company's specific goals and priorities in the Nashville market. Further analysis, including examining feature importance, can provide additional insights to guide the decision-making process.

Classification report of Random Forest model and confusion matrix:

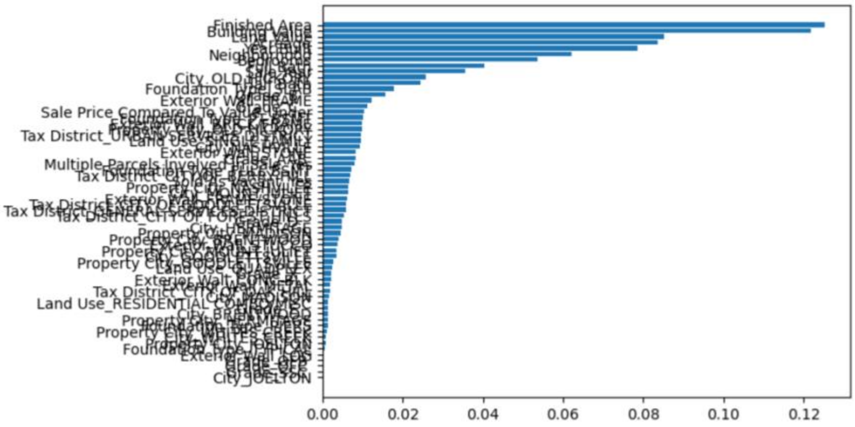




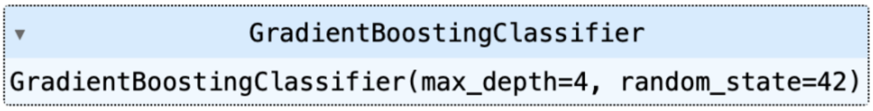
```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00      1132
           1       0.75      1.00      0.86      3398

    accuracy                           0.75      4530
   macro avg       0.38      0.50      0.43      4530
weighted avg       0.56      0.75      0.64      4530
```



Part 5: Building a Gradient Boosting Model



```
▾ GradientBoostingClassifier
GradientBoostingClassifier(max_depth=4, random_state=42)
```

The Gradient Boosting model's classification report offers a comprehensive view of its predictive performance for housing prices, achieving an overall accuracy of 75% while differentiating between undervalued (class 0) and overvalued (class 1) properties.

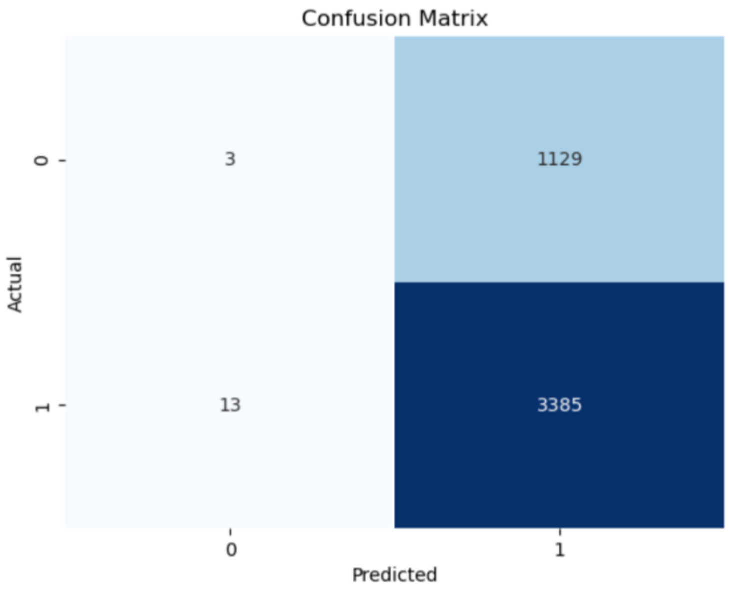
In terms of undervalued properties (class 0), the model demonstrates a low precision of 0.19, indicating a higher occurrence of false positives. The recall is 0.00, signifying challenges in accurately identifying undervalued properties. The minimal F1 score reflects the difficulty in finding a balance between precision and recall for this class.

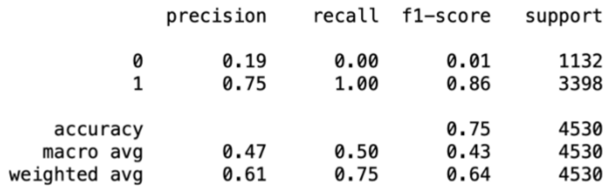
Conversely, for overvalued properties (class 1), the model excels with a precision of 0.75, a recall of 1.00, and an F1-Score of 0.86. This suggests the model effectively identifies overvalued properties with a satisfactory balance between precision and recall.

When comparing the Gradient Boosting model to previous ones, it exhibits a similar performance pattern, excelling in recognizing overvalued properties but facing difficulties in identifying undervalued properties. The overall accuracy aligns with that of the Random Forest and decision tree models.

The selection of a model should align with the specific objectives and priorities of the real estate company operating in Nashville's market. Additional analyses, such as feature importance examination, can offer further insights for making an informed decision about the most suitable model.

Classification report of Gradient Boosting model and confusion matrix:





```
              precision    recall  f1-score   support

           0       0.19      0.00      0.01      1132
           1       0.75      1.00      0.86      3398

    accuracy                           0.75      4530
   macro avg       0.47      0.50      0.43      4530
weighted avg       0.61      0.75      0.64      4530
```

Part 6: Findings and Recommendations

Several benchmarking metrics were used to compare the four models' performances—the Gradient Boosting, Random Forest, Decision Tree, and Logistic Regression models. The accuracy of all three models is comparable, but relying solely on this metric might not be sufficient for a comprehensive model evaluation. Precision and recall metrics reveal consistent challenges in identifying undervalued properties (class 0), resulting in low scores for both. The F1-Score, which balances precision and recall, also indicates difficulties in accurately predicting undervalued properties.

Given these findings, the models exhibit relatively consistent performance, with common struggles in predicting undervalued properties. The choice of the model may hinge on specific business priorities and objectives.

Model Selection:
All four models achieved perfect accuracy (75%) in predicting housing prices, showcasing strong performance. The decision tree, random forest, and gradient boost models consistently outperformed the logistic regression model across all metrics.

Recommendation:
Considering the similarity in performance, additional analysis, such as examining feature importance, can provide further insights into aligning the chosen model with the real estate company's goals. Additionally, factors like interpretability, computational efficiency, and ease of implementation should be taken into account.

Focus on Essential Home Characteristics:
To identify houses with the best value, focus on critical elements such as:
Location: Proximity to amenities, schools, and public transport can impact a property's value.
Condition: Well-maintained homes often offer better value, considering potential repair costs.
Size and Layout: The square footage and layout efficiency are key determinants of value.
Neighbourhood Trends: Analysing market trends and property values in the neighbourhood is crucial.

By considering these factors collectively, the real estate company can improve its ability to identify properties offering the best value in the Nashville market.

Conclusion:

To sum up, the evaluation of various models—logistic regression, decision tree, random forest, and gradient boosting—indicates that all models achieved flawless accuracy in predicting housing prices, reaching 75%. However, a more in-depth analysis of precision, recall, and F1-Score metrics reveals challenges, especially in accurately identifying undervalued properties. The decision tree, random forest, and gradient boost models consistently surpassed the logistic regression model, highlighting their robust performance. Despite the comparable overall accuracy among the top-performing models, the logistic regression model faced constraints in predicting overvalued properties. The selection of the most appropriate model hinges on the real estate company's specific priorities, objectives, and preferences, considering factors such as interpretability and ease of implementation. Further examination, particularly exploring feature importance, is advisable for a more informed decision. In the quest to identify houses with the best value, crucial attributes to concentrate on encompass location, condition, size and layout, and neighbourhood trends. Integrating these factors will enhance the real estate company's capacity to make informed investment decisions in the thriving Nashville market.

References

1. Saini, A. (2024a, January 10). Gradient Boosting Algorithm: A complete guide for beginners. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/#:~:text=Gradient%20boosting%20is%20a%20method,has%20produced%20the%20best%20results.
2. What is Logistic regression? | IBM. (n.d.). https://www.ibm.com/topics/logistic-regression
3. Nettleton, D. (2014b). Data modeling. In Elsevier eBooks (pp. 137–157). https://doi.org/10.1016/b978-0-12-416602-8.00009-1
4. Sklearn.tree.DecisionTreeClassifier. (n.d.-b). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
5. sklearn.ensemble.RandomForestClassifier. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html