***How does Random Forest regression model work?***

Random forest is a method used in modeling predictions and behavior analysis and is built on decision trees. It contains many decision trees representing a distinct instance of the classification of data input into the random forest. The Random Forest technique considers the instances individually, taking the one with the majority of votes as the selected prediction. In short, random forests are created from subsets of data and the final output is based on average or majority ranking.

Diagram 1 – Random forest process

Diagram, schematic

Description automatically generated

Note diagram 1 was referenced from (The Click Reader, 2022)

Random forest uses decision trees. A good example of decision tree was pointed out in our Bootcamp, Model 17.7.1 Overview of Decision Trees. The model used the example of a coin flip, whether it heads or tails is essentially how a decision tree uses decision analysis. “Decision tree convert a series of true/false questions that are represented by a seriers of if/else statements. Decision trees will help classify or label objects by asking a series of questions intended to zero in on the true answer.” (Arizona, 2022) Below are the key terms of a decision tree. To help illustrate see two diagrams.

Key terms: decision tree.

* **Root node/Parent node:** The top of the decision tree. Represents the entire population
* **Branch:**A subsection of the tree. Example of answering the “Yes” or “No” to a question.
* **Splitting:**The process of dividing a node into multiple sub-nodes.
* **Child node:** split the root node into two sub-notes. Example the two questions: “gone more than one week month?
* **Decision node:** When a sub-node is further split into additional sub-nodes.
* **Leaf node:** When a sub-node does not further split into additional sub-nodes; represents possible outcomes.
* **Pruning:** The process of removing sub-nodes of a decision tree. (Arizona, 2022)

Diagram 2 – Decision tree concept Diagram 3 – Decision Tree Process

Diagram

Description automatically generatedDiagram

Description automatically generated

Note diagram 2 was referenced from Model 17.7.1 Decision Trees. (Arizona, 2022) Note diagram 3 was referenced from (The Click Reader, 2022)

**Preprocessing that was involved:**

1. Load data in Pandas.

Text

Description automatically generated

Before dropping the columns in step 2, we took time to verify our categorical variable list and any unique values within each column. Looking at the dtypes(), meaning the type of data (integer, float, object...).

1. Drop columns that aren't useful



We dropped one column

1. Any missing data, verify?

Text

Description automatically generated

1. Define features set and target set
2. Create dummy variables

A picture containing table

Description automatically generated

Pandas get\_dummies has created a new 0/1 variable for every level of the original categorical variable. Pandas performed the operation on the columns that exists inside of the dataframe.

1. Divide the data set into training data and test data.

Graphical user interface, text, application, email

Description automatically generated

***Why this specific model?***

* It can do both regression and classification responsibilities.
* random forest produces good predictions that one is able to easily understood
* the ability to handle large datasets efficiently.
* many features can be performed at once

***What is the model's accuracy?***

RMSE is large indicating that our model is failing to account for important features underlying our data.

Text

Description automatically generated

First run through of regression

Text

Description automatically generated

Second run through of the regression and made modification to max\_depth.

***If there are statistics involved, what stats are being included in analysis and why?***

There was no statistics involved with this model. The use of RMSE (Root Mean Square Error). The RMSE indicates how close the actual data points are to the model’s predicted values. When RMSE value is low it indicates a better fit and is a good measure for determining the accuracy of the model’s predictions.

* ***If no statistics are involved, what would you include if you had more time?***

*No statistics were involved, used RMSE (Root Mean Square Error) because we are using regression model.*

# Works Cited

Arizona, U. o. (2022, August). *17.7.1 Overview of Decision Trees*. Retrieved from Bootcamp Spot: https://courses.bootcampspot.com/courses/2284/pages/17-dot-7-1-overview-of-decision-trees?module\_item\_id=723155

Team, C. (2022, December 20). *What is Random Forest?* Retrieved from Corporate Finance Institute: https://corporatefinanceinstitute.com/resources/data-science/random-forest/

The Click Reader. (2022). *TheClickReader*. Retrieved from Decision Tree Regression : https://www.theclickreader.com/decision-tree-regression/