

Business Analytics Final Project Task 1

MGT 7104

Yacine Ndiaye

May 5th,2025

Instructor: Jaime Velasco Sanchez

Richmond The American University in London

Word Count: 987

TABLE OF CONTENTS

TASK 1..... 2

INTRODUCTION..... 2

TASK1.1 3

TASK 1.2 3

TASK 1.3 5

CONCLUSION..... 5

TASK 1

INTRODUCTION

CD is a California based real estate company that specializes in residential real estate services. To further educate its licensed realtors on the experience of local markets, CD real estate wants to develop an analytical tool to predict the value of real estate.

This task will be using a CSV file called CALIREAL, it contains data on some census tract in California. The dataset counts 16,211 observations of 10 variables. To help CD real estate in their endeavour, this paper will run an analysis on the given dataset to predict the median house value using an individual regression tree after which the dataset will be split in 50% training and testing sets.

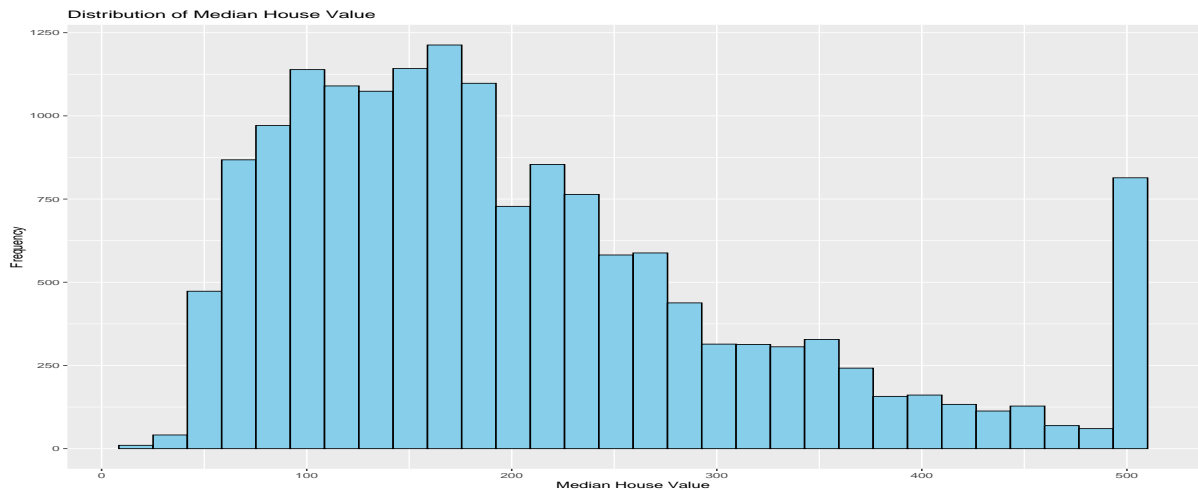
The variable index tract will not be included in this analysis. Other operations such as training a full and pruned regression tree to interpret and compare their RMSEs; computing the RMSE of the pruned tree on the test set and finally figuring out the predicted median house value for a specific tract using the settings from the pruned tree will be done. The packages used in the process are:

- “Rpart” and “Rpart.plot”: the former is the foundation for building decision tree models in R. it uses recursive partitioning to divide the data into increasingly homogenous groups based on input variables. (Milborrow,2021) the latter is a visualization package specifically designed for rendering decision trees generated by R part.
- “Caret”: this package is a comprehensive framework for training and evaluating machine learning models. In this context, it is used to split the data into training and testing sets, ensuring balanced and randomized partitions (Kuhn,2019)
- “Metrics”: offers a range of evaluation tools for assessing predictive models. In regression analysis, this package is particularly useful for calculating the RMSE, which quantifies the average magnitude of error between predicted and actual values. (Frasco, 2018)
- “Ggplot 2”: this package enables the creation of complex and customizable plots. It will be used to visualize distributions of median house values, relationship between features, model residuals and RMSE comparisons. (Pedersen,2024)

All elements involved in carrying out the aforementioned operations will be listed and explained along with supporting visualisations for interpretative ease.

TASK1.1

This Histogram depicts the distribution of the data after it's been split 50/50 for training and testing. The X axis represents median house value, it ranges from 0 to 500. The Y-axis represents frequency, it indicates how often house values in each range occurs within the dataset; the range goes from 0 to 1250. The graph shows that many houses are concentrated in the lower value bins (0-300) while houses valued between 300-500 are the lowest of the lot; there is a considerable peak for houses valued at 500.



TASK 1.2

This regression tree was used to predict median house value, based on variables such as median income, latitude, housing median age etc. The root node splits the data into two branches, left for lower income areas and right for higher income areas. Median income is < 4.5 . each node represents a decision rule which segments the data into increasingly smaller and more homogeneous groups (e.g., latitude ≥ 34) each yes/no branch corresponds to whether the condition is met. The leaf/terminal nodes are where predictions are made; each leaf contains a number and a percentage reflective of the observations from the path that falls in the node.

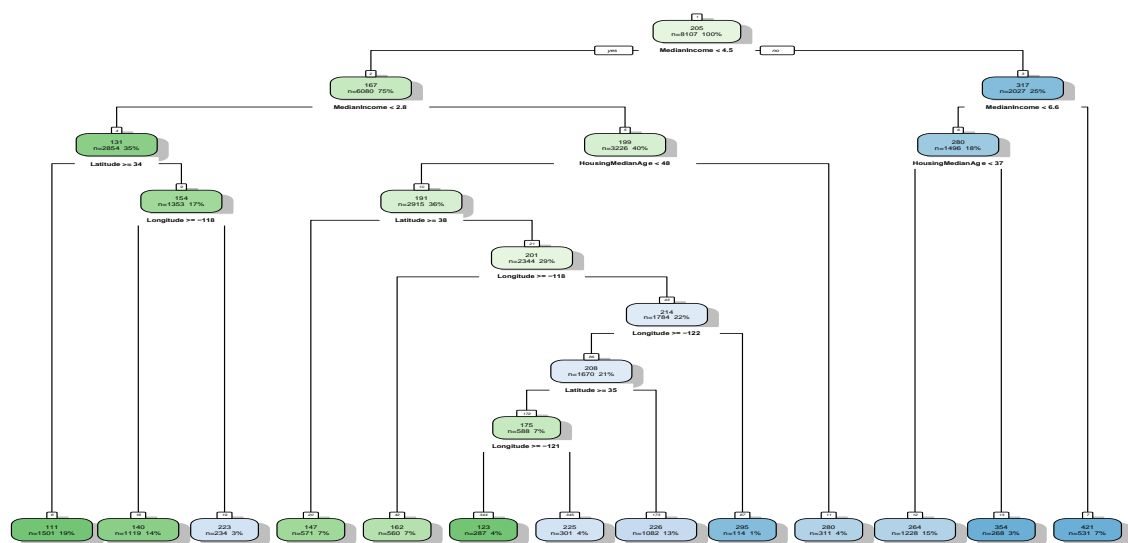
For instance, if we take the path Median income < 4.5 to Median income < 2.8 to latitude ≥ 34 to Longitude ≥ -118 . An interpretation of this path would be that in areas with median income below 2.8, latitude above 34 and longitude above -188, the predicted median house value is 131 as 35% of the observations from this path fall in this node.

The RMSE is 77.22 this equals the average prediction error when the full regression tree was tested on the validation dataset. Considering that house values range from 15.0 to 500.0, the number is moderate. This reflects considerably good accuracy.

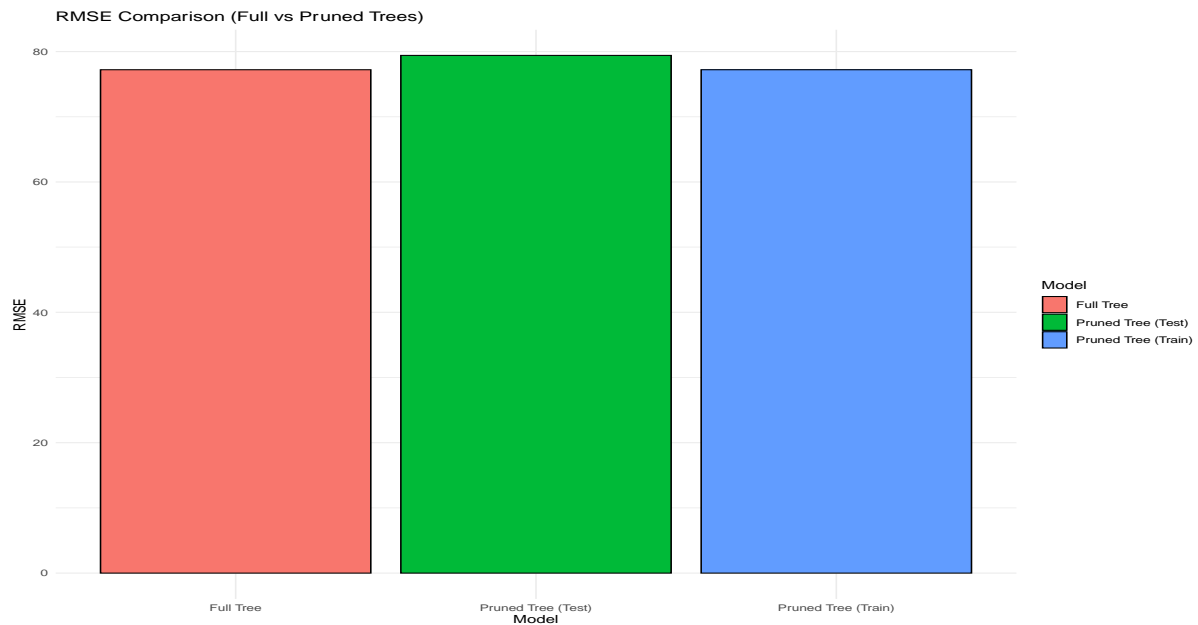
[illegible]

Median income is the most influential variable; geographic features (latitude, longitude) are used to segment the state spatially; smaller housing ages and higher income levels lead to higher house value predictions.

Pruned Regression Tree



Although the RMSE of the full and pruned regression trees from the training sets (77.22) bear no differences, the RMSE of the pruned tree from the testing set is slightly higher than the former (79.41). This small difference may be due to the pruned model being slightly less complex and underfitting certain patterns in unseen data. However, this trade-off helps improve model generalization and avoids overfitting.



TASK 1.3

For a tract that has a Longitude = -117.9, Latitude = 33.64, Age = 36, Rooms = 2, 107, Bedrooms = 357, Population = 850, Households = 348, and Income = 5.0532. the predicted median house value based on the pruned tree settings is 263.6. it's a moderately high amount.

CONCLUSION

While the model performs well, there is still room for improvements. Incorporating additional variables or testing alternative algorithms to improve accuracy for instance. Nonetheless, this project provides a robust foundation for CD Real Estate to make data-driven decisions in California's residential market. The use of regression trees offers a balance between interpretability and predictive power, making it a valuable tool for real estate analytics.

REFERENCES

- Frasco, M., 2018. "Evaluation metrics for machine learning". *RDocumentation*. Available at: <https://www.rdocumentation.org/packages/Metrics/versions/0.1.4> . Accessed: 01/05/2025
- Kuhn, M., 2019. "The caret package". *Topepo*. Available at: <https://topepo.github.io/caret/> . Accessed: 01/05/2025
- Milborrow, S., 2021. "Plotting Rpart trees with the Rpart. Plot package". Available at: <http://www.milbo.org/rpart-plot/prp.pdf> . Accessed: 01/05/2025
- Pedersen, T.L., 2024. "GGPLOT2". *RDocumentation*. Available at: <https://www.rdocumentation.org/packages/ggplot2/versions/3.5.0> . Accessed: 01/05/2025

INDEX

RMSE:Root Mean Squared Error

MSE:Mean Square Error

CP:Complexity Parameter

R-CODE TASK 1

STEP 1: LOAD DATA (Q1)

```
CL <- read.csv("CL.CSV")
```

```
CL$Tract <- NULL # Remove index variable as instructed
```

```
CL <- na.omit(CL)
```

```
CL$MedianHouseValue <- as.numeric(CL$MedianHouseValue)
```

Install and load required packages

```
required_packages <- c("rpart", "rpart.plot", "caret", "Metrics", "ggplot2")
```

```
for (pkg in required_packages) {
```

```
  if (!require(pkg, character.only = TRUE)) {
```

```
    install.packages(pkg, dependencies = TRUE)
```

```
    library(pkg, character.only = TRUE)
```

```
  }
```

```
}
```

Visualize target variable

```
ggplot(CL, aes(x = MedianHouseValue)) +
```

```
  geom_histogram(bins = 30, fill = 'skyblue', color = 'black') +
```

```
  labs(title = "Distribution of Median House Value", x = "Median House Value", y =  
"Frequency")
```

STEP 2: SPLIT DATA (50% TRAIN, 50% TEST) (Q1)

```
set.seed(123)
```

```
train_index <- createDataPartition(CL$MedianHouseValue, p = 0.5, list = FALSE)
```

```
train_data <- CL[train_index, ]
```

```
test_data <- CL[-train_index, ]
```

STEP 3: TRAIN FULL REGRESSION TREE (Q2, Q3)

```
full_tree <- rpart(MedianHouseValue ~ ., data = train_data, method = "anova")
```

```
rpart.plot(full_tree, main = "Full Regression Tree")
```

```
# Evaluate on training data (Validation Experiment) (Q2/Q3)
train_pred_full <- predict(full_tree, train_data)
rmse_full_train <- rmse(train_data$MedianHouseValue, train_pred_full)
cat("Q2/Q3 - RMSE (Full Tree, Training):", round(rmse_full_train, 2), "\n")
```

```
# STEP 4: PRUNE TREE USING BEST CP (Q4, Q5)
```

```
printcp(full_tree)
best_cp <- full_tree$sctable[which.min(full_tree$sctable[, "xerror"]), "CP"]
pruned_tree <- prune(full_tree, cp = best_cp)
```

```
rpart.plot(pruned_tree, main = "Pruned Regression Tree", type = 2, extra = 101,
           box.palette = "GnBu", shadow.col = "gray", nn = TRUE)
```

```
# Evaluate pruned tree on training data (Validation Experiment) (Q4/Q5)
```

```
train_pred_pruned <- predict(pruned_tree, train_data)
rmse_pruned_train <- rmse(train_data$MedianHouseValue, train_pred_pruned)
cat("Q4/Q5 - RMSE (Pruned Tree, Training):", round(rmse_pruned_train, 2), "\n")
```

```
# STEP 5: COMPARE RMSEs (Q6)
```

```
cat("Q6 - RMSE Comparison:\n")
cat(" Full Tree RMSE:", round(rmse_full_train, 2), "\n")
cat(" Pruned Tree RMSE:", round(rmse_pruned_train, 2), "\n")
```

```
# STEP 6: EVALUATE PRUNED TREE ON TEST SET (Q7, Q8)
test_pred_pruned <- predict(pruned_tree, test_data)
rmse_pruned_test <- rmse(test_data$MedianHouseValue, test_pred_pruned)
cat("Q7/Q8 - RMSE (Pruned Tree, Test Set):", round(rmse_pruned_test, 2), "\n")
```

```
# STEP 7: VISUALIZE PERFORMANCE
# RMSE comparison between full and pruned trees
rmse_comparison <- data.frame(
  Model = c("Full Tree", "Pruned Tree (Train)", "Pruned Tree (Test)"),
  RMSE = c(rmse_full_train, rmse_pruned_train, rmse_pruned_test)
)
ggplot(rmse_comparison, aes(x = Model, y = RMSE, fill = Model)) +
  geom_bar(stat = "identity", color = "black") +
  labs(title = "RMSE Comparison (Full vs Pruned Trees)", y = "RMSE") +
  theme_minimal()
```

```
# STEP 8: PREDICT MEDIAN VALUE FOR SPECIFIC TRACT (Q9)
new_tract <- data.frame(
  Longitude = -117.9,
  Latitude = 33.64,
  HousingMedianAge = 36,
  TotalRooms = 2107,
  TotalBedrooms = 357,
  Population = 850,
  Households = 348,
  MedianIncome = 5.0532
)
```

```
pred_value <- predict(pruned_tree, new_tract)
cat("Q9 - Predicted Median House Value for New Tract:", round(pred_value, 2), "\n")
```