Essay: Decision Trees in Car Purchase Prediction and Addressing Common Criticisms

# 1. Introduction to Decision Trees for Classification

Decision trees are widely used in machine learning for their simplicity, interpretability, and flexibility in handling classification tasks. For this project, I selected a very simple dataset aimed at predicting car purchases based on basic demographic variables: age, gender, and salary. This dataset was chosen to make decision trees manageable and to clearly observe how model performance and interpretability vary under different configurations. In practice, more complex datasets would likely require additional preprocessing and model tuning, as deeper, more intricate trees would be necessary. Even with this dataset, however, the default decision tree depth was deeper than what our optimized grid search identified as optimal.

This analysis involved constructing three decision trees and a random forest model to investigate the effects of variable selection and model tuning. Each model had a unique configuration: a default decision tree, a decision tree with grid-searched optimal depth, a randomized decision tree with a forced initial split, and a random forest. Comparing these models allowed me to assess how tuning and randomization influence decision tree performance, interpretability, and stability.

# 2. Data Preparation and Exploratory Analysis

The dataset used includes demographic features—age, gender, and salary—along with a target variable indicating whether a car was purchased. Before model construction, I conducted a thorough exploratory data analysis (EDA), provided in an attached Jupyter notebook. The EDA included several plots and insights, which highlighted patterns in the data and informed our approach to feature selection.

## EDA Visualizations

A screenshot of a graph

Description automatically generated

1. **Density and Count Plot Grid:**
   * **Age vs. Purchased Density Plot:** This plot displays the age distribution among individuals who purchased a car versus those who didn’t. The distribution suggests that younger individuals are less likely to purchase a car, with the likelihood increasing moderately with age.
   * **Salary vs. Purchased Density Plot**: A clear separation is observed between the salary distributions of those who purchased a car and those who didn’t, with higher salaries correlating strongly with a greater likelihood of purchase. This finding aligns with our models, which identified salary as an important predictor of car buying behavior.
   * **Purchase Decision by Gender:** This count plot reveals only a slight difference in purchasing behavior between genders, supporting the low importance assigned to gender by the models.
   * **Purchased Distribution:** This plot shows that the dataset is slightly imbalanced, with more individuals not purchasing a car, which could potentially affect the model’s sensitivity to each class.

A diagram of a heatmap

Description automatically generated

1. **Correlation Heatmap:** The correlation heatmap provides a visual of the relationships between age, salary, gender, and car purchase decision. Both age and salary show moderate correlations with the target variable, indicating that they are valuable predictors, whereas gender has almost no correlation with purchase behavior, a finding echoed by the low feature importance for gender in all models.

# 3. Model Building and Performance Comparison

To explore the effects of feature importance and model tuning, I constructed four models:

* **Decision Tree Auto:** This default decision tree used scikit-learn's default settings. With no tuning, the tree went deeper than necessary, increasing its risk of overfitting and making it vary messy and overwhelming to look at.

A diagram of a network

Description automatically generated

* **Decision Tree 1 (Optimized):** Using grid search, I found the optimal max depth and min samples per split, achieving a well-balanced model with good performance and minimized overfitting.

A diagram of a diagram

Description automatically generated

* **Decision Tree 2 (Randomized Split):** This model used a forced random initial split to investigate the effect of initial splitting decisions on performance and interpretability.

A diagram of a company

Description automatically generated

* **Random Forest:** This ensemble model, created with grid search for parameter optimization, combines multiple decision trees to reduce variance and improve stability.

The feature importances of these models reveal consistent trends for age and salary, with salary being particularly important in Decision Tree 2. Here is a comparison:

**A screenshot of a graph

Description automatically generated**

Age and salary consistently emerged as key predictors, with salary’s importance being higher in Decision Tree 2 and slightly lower but stable in the Random Forest. Gender was minimally influential across all models, aligning with the insights from our EDA.

# Model Performance Comparison

Each model’s performance metrics—accuracy, F1 score, and ROC AUC—were also compared to determine the most effective model. The results are as follows:

**A screenshot of a graph

Description automatically generated**

The Random Forest and Decision Tree 1 (Optimized) demonstrated the best overall performance. Decision Tree Auto showed overfitting tendencies, while Decision Tree 2’s randomized split led to significant variance and instability. This highlights the importance of parameter tuning, as seen with Decision Tree 1 and the Random Forest, which delivered higher accuracy and better balance across metrics.

# 4. Addressing “The Good, The Bad, and The Ugly” of Decision Trees

In line with the blog post from Decizone, our findings demonstrate several key strengths and limitations of decision trees:

* **Overfitting:** The default decision tree (Decision Tree Auto) showed signs of overfitting, with a greater depth than necessary. To address this, I optimized tree depth and split requirements using GridSearchCV, which proved effective in Decision Tree 1. The Random Forest further mitigated overfitting through ensembling.
* **Interpretability vs. Complexity:** Decision trees are highly interpretable at shallow depths. With additional features, however, interpretability could become a challenge. The optimized Decision Tree 1 struck a good balance between interpretability and performance.
* **Bias and Variance Trade-Off**: Decision trees are sensitive to initial splits, and random splitting in Decision Tree 2 led to notable performance degradation. The Random Forest, by averaging results across multiple trees, helped reduce variance without increasing bias.
* **Data Imbalance:** Imbalance in the target classes was a factor in our analysis, as seen in the purchase distribution plot. The accuracy and ROC AUC metrics used partially offset this issue, though further work with class weights or resampling could enhance the model's handling of minority class predictions.

# 5. Conclusion and Real-World Findings

This analysis provided dual insights: it demonstrated both the practical application of decision trees in classification and revealed meaningful predictors of car purchases in a simple, real-world dataset.

From a modeling perspective, decision trees proved useful for interpreting variable importance, and the random forest’s ensemble approach delivered the most stable and high-performing results. While the default decision tree classifier was overly deep, likely contributing to overfitting, the optimized decision tree (DT1) achieved high accuracy and interpretability. In contrast, the decision tree with a random initial split (DT2) exhibited high variance, reinforcing that tree stability depends on robust parameter optimization. Using grid search, both the random forest and DT1 achieved optimized depth and balanced accuracy, F1, and ROC AUC scores, showing that fine-tuning improves decision trees’ real-world application by mitigating overfitting risks.

On the practical side, this analysis illuminated key drivers behind car-buying behavior. Age and salary emerged as significant predictors, with older and higher-income individuals showing a stronger likelihood to purchase cars. This aligns with broader consumer trends: younger people, especially those in lower-income brackets, may delay car purchases due to cost constraints or lifestyle factors. Age was the most impactful feature across models, underscoring its importance as a predictor, while salary also had a substantial impact. Gender, however, contributed minimal predictive power, suggesting it plays a lesser role in car-buying decisions.

This outcome suggests that car dealerships and automotive marketers may achieve better results by targeting promotions and financing options based on age and income rather than gender. The clear relationships observed between age, salary, and car purchase decisions in this straightforward dataset affirm that decision trees can reveal meaningful insights into customer behavior when applied thoughtfully. While this dataset was relatively simple, future applications on more complex data would likely benefit from additional tuning and deeper analysis to prevent overfitting and ensure model reliability.