DATA MINING AND MACHINE LEARNING

ASSIGNMENT 1 : Classification

March 17, 2024

Submitted by

Hiba A P [MDS202326]

Nooh Ali [MDS202337]

INTRODUCTION

We are provided with two data sets and based on that we are applying a supervised machine learning method where the model aims to predict the correct label or category for a given input data.

Task 1: Customer Churn
The customer churn dataset captures interactions with an online retail store. We Built two classifiers to predict churn: one using Adaboost and one using random forest. Using an evaluation metric we are comparing their performance.

Task 2: Supermarket Sales
The supermarket sales dataset contains sales data for a three-branch supermarket. We built: Gender Predictor and Rating Predictor using classifiers (one with a decision tree and another with a random forest) and using an evaluation metric we are comparing their performance.

REPORT

Our Achievement

Task 1

Accuracy of Classifier for the customer churn in online retail store

Data \ Method Used	Random Forest	ADAboost	
		Decision Tree	Logistic Regression
Training Data	61%	84.3%	54.3%
Testing Data	59.3%	54%	55.2%

Comparison

In this case our data is highly scattered and noisy, ADAboost is highly sensitive than Random Forest to noisy data. So, Random Forest gives better predictions.
Random Forest is not over-fitted compared to ADAboosting (decision tree).

The random forest model achieves a reasonable balance between training and testing accuracy. While the testing accuracy is slightly lower than the training accuracy, it indicates that the model is not significantly overfitting. Random forests are robust and versatile, making them suitable for various datasets.

The AdaBoost decision tree model shows a substantial gap between training and testing accuracy. The high training accuracy suggests overfitting, as the model performs exceptionally well on the training data but struggles to generalise to unseen data. AdaBoost focuses on combining weak learners (stumps), but in this case, it may not be achieving the desired generalisation.

The linear regression-based AdaBoost model demonstrates similar performance on both training and testing data. While it avoids severe overfitting, it doesn’t outperform the other models significantly. Linear regression as a base model may not capture complex relationships effectively.

Task 2

Accuracy of Gender Predictor of the Supermarket data

Data \ Method Used	Random Forest	Decision Tree	Logistic Regression
Training Data	97%	54.9%	57.4%
Testing Data	54%	52.3%	50%

Comparison

The Decision Tree model seems to strike a better balance between training and testing accuracies, making it the preferred choice. The overfitting occurs when a model captures noise and fluctuations in the training data, leading to poor generalisation on new data. Therefore, we aim for models that generalise well without being overly complex or simplistic.

The Random Forest model is likely overfitted due to the significant difference in accuracies. The Decision Tree model seems to be less overfitted compared to the Random Forest. However, it still exhibits some overfitting due to the modest training accuracy. The Linear Regression model appears to be less overfitted than the Random Forest but still lacks generalisation power. It might benefit from more feature engineering or regularisation techniques.

Error in the Rating Predictor of the Supermarket data

Data \ Method Used	Linear Regression		Decision Tree Regression	
	MSE	R Square	MSE	R Square
Training Data	0.2939	0	2.907	0.0107
Testing Data	0.297	-0.0029	2.998	-0.0092

Comparison

The Linear Regression model is not overfitted, but its predictive power is limited. The negative R-squared error for testing data indicates that the model lacks explanatory capability. ive R-squared error for testing data indicates that the model lacks explanatory capability.

The MSE for both testing and training data is relatively low, indicating that the model’s predictions are close to the actual values. However, the negative R-squared error for testing data is concerning. It suggests that the model performs poorly in explaining the variation in the response variable. The zero R-squared error for training data implies that the model doesn’t explain any variation in the training data.

The MSE for both testing and training data is higher than that of Linear Regression, suggesting more prediction errors. The negative R-squared error for testing data indicates poor explanatory power. The positive R-squared error for training data suggests some ability to explain variation.

What we did?

Exploratory Data Analysis

To understand the structure of the data we've plotted graphs like

1. Barplot
2. Scatter Plot
3. Pair Grid Plot
4. Violin Plot
5. Correlation Heatmap
6. Swarm Plot
7. Pie Plot

Classification Problem

1. Checked for Null Values: Ensured data completeness by checking for missing values in the dataset.

2. Scaled the Values for Logistic Regression: Performed feature scaling to standardize or normalize numerical features specifically for logistic regression, considering its sensitivity to feature scales.

3. Tried One-Hot Encoding for Categorical Variables: Initially attempted one-hot encoding to convert categorical variables into a numerical format. However, found it didn't improve accuracy, so opted for label encoding instead.

4. Fitted a Random Forest Classifier: Employed the Random Forest algorithm, a robust ensemble learning technique based on decision trees, suitable for capturing complex feature interactions and achieving high accuracy in predicting customer churn.

5. Applied AdaBoost on Decision Tree: Utilized AdaBoost, an ensemble method enhancing decision trees' performance by iteratively training weak learners on subsets of data, adjusting weights of misclassified samples to improve predictive performance.

6. Applied Logistic Regression: Utilized logistic regression, a binary classification algorithm modelling the probability of customer churn based on predictor variables. Logistic regression provides interpretable results and is commonly used for binary classification tasks.

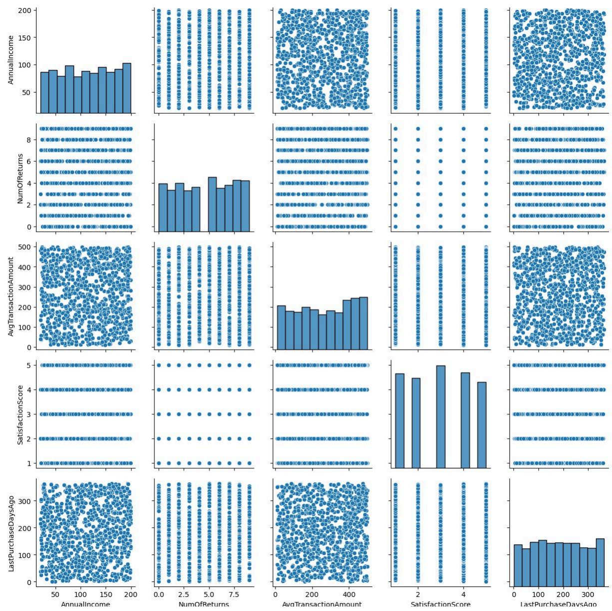
Regression Prediction

1. Fitted Decision Tree Regressor and Linear Regression: Implemented both Decision Tree Regressor and Linear Regression algorithms to predict the rating of customers. These algorithms were chosen due to their suitability for regression tasks and ability to model the relationship between features and the target variable.

2. Utilized Recursive Feature Elimination (RFE) for feature selection in the linear regression model.

3. Employed Grid Search for hyperparameter tuning to optimize the performance of the linear regression model.

Our Challenges



Scattered Data Distribution:

The Pair Grid plot revealed that the given data points were widely dispersed, lacking any discernible pattern. The scattered nature of the data posed difficulties for model fitting. Without clear trends or relationships, achieving high accuracy became a formidable task.

Lack of Predictive Patterns:

Despite rigorous modeling efforts, all our fitted models yielded disappointingly low accuracies. Explanation: The absence of consistent predictive patterns hindered our ability to build effective models. Without identifiable features driving the outcomes, our models struggled to generalize. We needed to explore alternative approaches beyond traditional feature-based modeling.

Feature Selection Challenges:

Attempting feature selection proved challenging due to the unique characteristics of our data. The data defied many of our logical assumptions. For instance, the number of purchases and the number of support queries exhibited no significant correlation. Without clear feature relevance, selecting an optimal subset for modeling became elusive.

Time Constraints:

Our project operated under tight time constraints. The limited timeframe restricted our ability to explore more sophisticated techniques or perform exhaustive hyperparameter tuning. Trade-off: While we achieved results within the deadline, we acknowledge that additional time could have led to more refined models.

Deep Learning Unfamiliarity:

Our team lacked familiarity with deep learning models. Deep learning techniques, such as neural networks, could potentially handle complex relationships in the data. We believe that if we had deeper expertise in these models, we might have achieved better performance.