Predicting Ratings from SuperMarket Sales Data

Utpalraj Kemprai, Aryan Joshi

March 2024

1 Introduction

Our goal is to build two models to predict Rating, one using linear regression and one using a decision tree regressor using the Supermarket sales dataset.

2 Exploratory Data Analysis

We carried out our analysis in the following manner:

1. We first checked the datatypes of all columns in the data.
2. We checked for missing values and found there were no missing values in the data
3. We plotted the correlation Heat map for numeric columns and found that Tax and Total had perfect correlation.
4. We checked for outliers by standardising all the numeric columns and then plotting boxplot diagrams for them. We found a few outliers in Tax and Total (some values lied outside outlier fences). Since these two are perfectly correlated, the outliers corresponded to the same rows in the dataset.
5. We then plotted the histogram to check the distribution of customer ratings. We found that the distribution was roughly uniform.

3 Data Preprocessing

- For preprocessing the data we encoded all the non-numeric columns using One-Hot encoder.
- We dropped the InvoiceID column because it was obviously independent of the Customer Ratings. We also dropped the Tax column because it had perfect correlation with the Total column.
- We split the data into train and test data (80% for Training and 20% for testing).

4 Training the models

4.1 Linear Regression

4.1.1 Approach

We fit the following models on the data:

1. Linear model(degree 1) without regularization
2. Linear model (degree 1) with Lasso
3. Linear model (degree 1) with Ridge
4. Linear model (degree 1) with Elastic Net
5. Quadratic polynomial with lasso regularisation
6. Quadratic polynomial with Ridge
7. Quadratic Poynomial with Elastic Net
8. Quadratic Poynomial without regularization

4.1.2 Results

We observed the following RMSE and R^2 values on train data:

Model	RMSE	R^2
Linear model(degree 1) without regularization	1.7216	0.0096
Linear model (degree 1) with Lasso	1.7297	0.0003
Linear model (degree 1) with Ridge	1.7216	0.0096
Linear model (degree 1) with Elastic Net	1.7278	0.0025
Quadratic polynomial with lasso regularisation	1.7212	0.0101
Quadratic polynomial with Ridge	1.6481	0.0924
Quadratic Poynomial with Elastic Net	1.7225	0.0086
Quadratic Poynomial without regularization	1.6480	0.0925

Table 1: Model Performance on Train set.

We observed the following RMSE and R^2 values on test data:

Model	RMSE	R^2
Linear model(degree 1) without regularization	1.6704	-0.0139
Linear model (degree 1) with Lasso	1.6581	0.0008
Linear model (degree 1) with Ridge	1.6694	-0.0127
Linear model (degree 1) with Elastic Net	1.6584	0.0005
Quadratic polynomial with lasso regularisation	1.6648	0.0101
Quadratic polynomial with Ridge	1.6977	-0.0474
Quadratic Poynomial with Elastic Net	1.6659	0.0084
Quadratic Poynomial without regularization	1.7011	-0.0516

Table 2: Model Performance on Test set.

4.2 Decision Tree Regressor

4.2.1 Approach

We performed hyperparameter tuning for the Decision Tree Regressor on the max_depth, max_features, strategy used to choose the split(splitter) and the cost complexity pruning parameter (ccp_alpha).

4.2.2 Results

- The best parameters found were max_depth = 3, max_features = 'sqrt', splitter = 'random' and ccp_alpha = 0.01.
- The RMSE score and R^2 on train data were 1.7170 and 0.0149 respectively.
- The RMSE score and R^2 on test data were 1.6577 and 0.0013 respectively.

5 Comparing the results

From the above results we see that Linear Regression Model of order 2 with L1 regularization has the highest R^2 among all the regression models we implemented for predicting the ratings on the test data. The lowest RMSE score was obtained by the Decision Tree Regressor and since we are building a model for predicting customer ratings, our primary goal is accurate predictions and so the Decision Tree Regressor proves be a better model than Linear regressor although by a small margin based on our results on the test data.