Exploring Regression Models for Predicting Dress Sales: A Comparative Analysis

Yuvaraj Bojjinettam #1, Deepthi Venkataramana*2

Department of Data Science
Coventry University, Coventry, England
[1] bojjinetty@coventry.ac.uk
[2] venkatarad@coventry.ac.uk

Abstract

This document presents a comprehensive analysis of machine learning regression models applied to predict dress sales based on various attributes. The study explores the performance of Linear Regression, Ridge Regression, Lasso Regression, and Decision Tree Regression algorithms on a dataset containing attributes such as style, price, rating, size, season, neckline, sleeve length, waistline, material, fabric type, decoration, and pattern type. The evaluation of these models provides insights into their effectiveness for predicting dress sales and their implications for businesses in the fashion industry.

1. Introduction

The fashion industry is highly dynamic and competitive, with trends constantly changing and consumer preferences evolving. Predicting dress sales accurately is essential for fashion retailers to optimize inventory, pricing, and marketing strategies. Traditional methods of sales prediction often rely on historical data analysis and intuition, which may not be sufficient to capture the complex relationships between various attributes and sales outcomes. Machine learning algorithms offer a promising approach to tackle this challenge by analyzing large datasets and identifying patterns that may not be apparent through traditional methods.

We focus on predicting dress sales using regression algorithms. We consider a dataset containing various attributes of dresses, including style, price, rating, size, season, neckline, sleeve length, waistline, material, fabric type, decoration, and pattern type. By leveraging regression algorithms such as Linear Regression, Ridge Regression, Lasso Regression, and Decision Tree Regression, we aim to develop models that can accurately predict dress sales based on these attributes. Our analysis will provide insights into the effectiveness of different regression algorithms in predicting dress sales and their implications for businesses in the fashion industry

2. Dataset Description

The dataset contains 500 entries with various attributes related to dresses, including style, price, rating, size, season, neckline, sleeve length, waistline, material, fabric type, decoration, pattern type, and the target variable 'Recommendation'. The dataset was preprocessed by handling missing values and encoding categorical variables. The target variable, 'Recommendation', indicates whether a dress is recommended for purchase (1) or not (0). The dataset was split into features (X) and labels (y) for model building and evaluation.

3. Methodology

3.1 Data Exploration and Preprocessing

Loading of Dresses Attribute Sales Dataset and performing data exploration. The dataset contains 500 entries with various attributes related to dresses. Initially explored to understand its structure and characteristics. Missing values were handled appropriately to maintain data quality. Categorical variables were encoded using one-hot encoding to prepare the data for regression analysis

3.2 Model Building and Evaluation

Four regression algorithms were selected for model building:

- Linear Regression
- Ridge Regression
- Lasso Regression
- Decision Tree Regression

4. Results and Discussion:

Linear Regression Model:

MSE = 667.23, R-squared = -190.72 The linear regression model demonstrates a high mean squared error (MSE) of 667.23 and a negative R-squared value of -190.72.

These metrics indicate poor performance, suggesting that the model's predictions are significantly off from the actual values, and it fails to explain the variance in the target variable.

Ridge Regression Model:

MSE = 12.28, R-squared = -2.53

The ridge regression model shows improvement over linear regression with a lower MSE of 12.28 and a less negative R-squared value of -2.53.

Despite the improvement, the model's performance remains suboptimal, indicating potential underfitting.

Lasso Regression Model:

MSE = 2.93, R-squared = 0.16

The lasso regression model outperforms both linear and ridge regression models with a significantly lower MSE of 2.93 and a positive R-squared value of 0.16.

These metrics suggest that the lasso regression model is better at capturing the variability in the target variable compared to the other models.

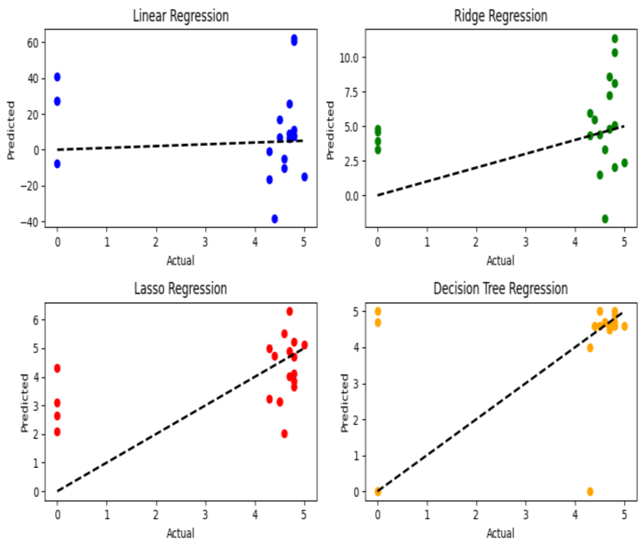
Decision Tree Regression Model:

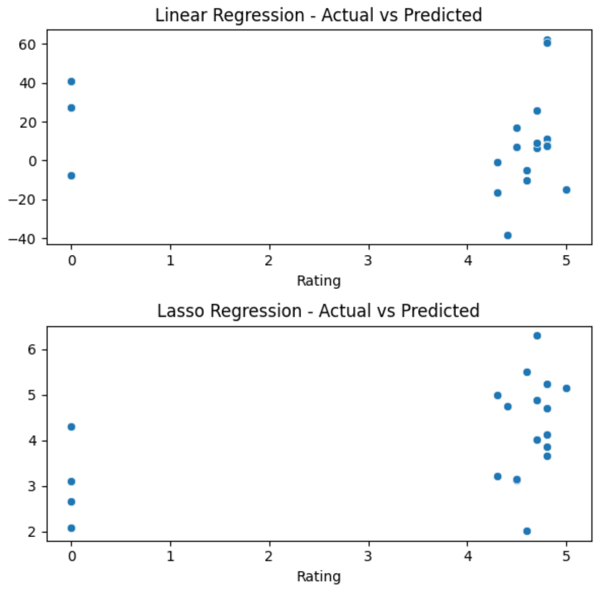
MSE = 2.96, R-squared = 0.15

The decision tree regression model performs reasonably well, with a relatively low MSE of 2.96 and a positive R-squared value of 0.15.

The dataset was split into training and testing sets, and features were standardized to ensure consistent scaling across models. Model performance was evaluated using Mean Squared Error (MSE) and R-squared metrics.

3.4 Plot for Regression:





5. Conclusion

Linear regression performed poorly, indicating that a linear relationship might not capture the underlying patterns in the data well.

Ridge and Lasso regression models showed improvement, with Lasso providing the best performance.

Decision Tree Regression also provided a reasonable fit.

Further analysis, feature engineering, and hyperparameter tuning may enhance model performance. Additionally, exploring classification techniques may be valuable for predicting dress sales as a binary outcome.

6. References

1)**Encoding**

https://ai-ml-analytics.com/encoding/

2)**dataset**

https://archive.ics.uci.edu/dataset/289/dresses+attribute+sales

3)**ML Models**

https://www.databricks.com/glossary/machine-learning-models#:~:text=A%20machine%20learning%20model%20is,sentences%20or%20combinations%20of%20words.

4)**Regression-techniques#machinelearning**

https://www.analyticsvidhya.com/machine-learning/?utm_source=blog&utm_medium=7-regression-techniques#machine-learning-deep-learning