Automated Feature Selection for Improving Predictive Model Robustness

Yagel Alankry & Ido Keren
Tabular Data Science Final Project

March 20, 2025

Abstract

Feature selection plays a crucial role in predictive modeling by improving model accuracy, reducing computational cost, and enhancing interpretability. In this project, we develop an **automated feature selection pipeline** using SHAP-based ranking, Recursive Feature Elimination (RFE), and correlation-based filtering. Our method was tested on four datasets, comparing its performance against a naive approach using all features. The results demonstrate improved model robustness, higher R^2 scores, and reduced overfitting. This project incorporates methodologies from *Guyon et al. (2003)* on variable selection and *Lundberg et al. (2017)* on SHAP values to create a structured and efficient feature selection framework.

1 Problem Description

Traditional feature selection techniques often require significant domain expertise and are time-consuming, especially when dealing with large datasets. Manual feature selection can become increasingly difficult as the feature set grows, leading to inefficiencies. This increases the importance of developing an automated feature selection pipeline that can handle large-scale datasets and dynamically select the most relevant features for the model. The goal of this project is to design and implement an automated feature selection pipeline that improves the model's performance while ensuring computational efficiency.

2 Solution Overview

The proposed solution is an automated feature selection pipeline that combines multiple techniques for feature filtering and ranking:

1. **Low-Correlation Feature Removal**: Features that show low correlation with the target variable are removed from the dataset to reduce noise.
2. **SHAP (Shapley Additive Explanations) Values**: SHAP values are used to evaluate the impact of each feature on the model's predictions. Features with higher SHAP values are considered more important for the model's accuracy.
3. **Recursive Feature Elimination (RFE):** This technique helps identify the subset of features that contribute the most to model performance by recursively eliminating the least important features.

The combination of these steps ensures that only the most relevant features are retained, improving model performance while reducing overfitting.

3 Implementation of Academic Articles

3.1 Guyon et al. (2003) - Feature Selection

This paper provided theoretical foundations for feature selection methods, emphasizing the importance of:

- Ranking features based on relevance and redundancy.
- Using both filter (correlation) and wrapper (SHAP) methods.
- Evaluating different subsets of features iteratively.

We implemented these principles by integrating correlation filtering and SHAP-based ranking to enhance interpretability and efficiency.

3.2 Lundberg et al. (2017) - SHAP Values

SHAP values provide a unified measure of feature importance using game-theoretic principles. We incorporated SHAP for:

- Ranking features by their contribution to model predictions.
- Selecting optimal features iteratively, improving interpretability.
- Comparing feature importance across multiple datasets.

4 Experimental Evaluation

We tested our pipeline on four datasets:

- NYC Airbnb Prices
- Amazon Bestselling Books
- Global Air Pollution Data
- Salary Prediction
- Sleep Efficiency Dataset

We compare the naive approach (using all features) to our automated method using XGBoost.

4.1 Performance Metrics

Table 1: Model Performance Comparison

Metric	**Naive Model**	**Automated Selection**
MAE	High	Low
MSE	High	Low
R^2 Score	Negative	Positive
Pearson Correlation	Low	High

4.2 Dataset 1: Airbnb New York (AB_NYC_2019)

The first dataset contains details about Airbnb listings in New York for the year 2019. It is used to predict the price of Airbnb listings based on various features such as location, room type, and availability.

Dataset Dimensions:

- **Number of Features:** 16
- **Number of Samples:** 48,000

Target Variable:

- **Price:** The price per night of the listing.

Results:

Metric	Naive Solution	Automated Solution
MAE	67.25	71.16
MSE	45058.17	40056.32
R^2	-0.0185	0.0945
Pearson	0.3712	0.3106

Table 2: Airbnb New York Dataset - Model Performance

Overfitting Prevention: Selecting Optimal Feature Count The plot below shows the relationship between the number of features used in the model and the resulting R^2 score. Initially, we see a small increase in the R^2 score as more features are included. However, beyond a certain point, the score begins to fluctuate, indicating that the model may be overfitting the data. Overfitting occurs when the model becomes too complex by including too many features, which leads to poor generalization on unseen data.

As the number of features increases from 1 to 9, the R^2 score fluctuates, indicating that additional features sometimes improve the model, but at other times, they lead to a decrease in performance. This suggests that using too many features may cause the model to overfit, especially if the features are noisy or irrelevant. To avoid overfitting, it is important to select a feature set that improves the model's performance without adding unnecessary complexity.

The fluctuations in the R^2 score reinforce the idea that there is an optimal number of features where the model performs best. In this case, it would be prudent to select a feature count where the R^2 score is relatively stable and not subject to drastic fluctuations, thereby ensuring the model remains both accurate and generalizable.

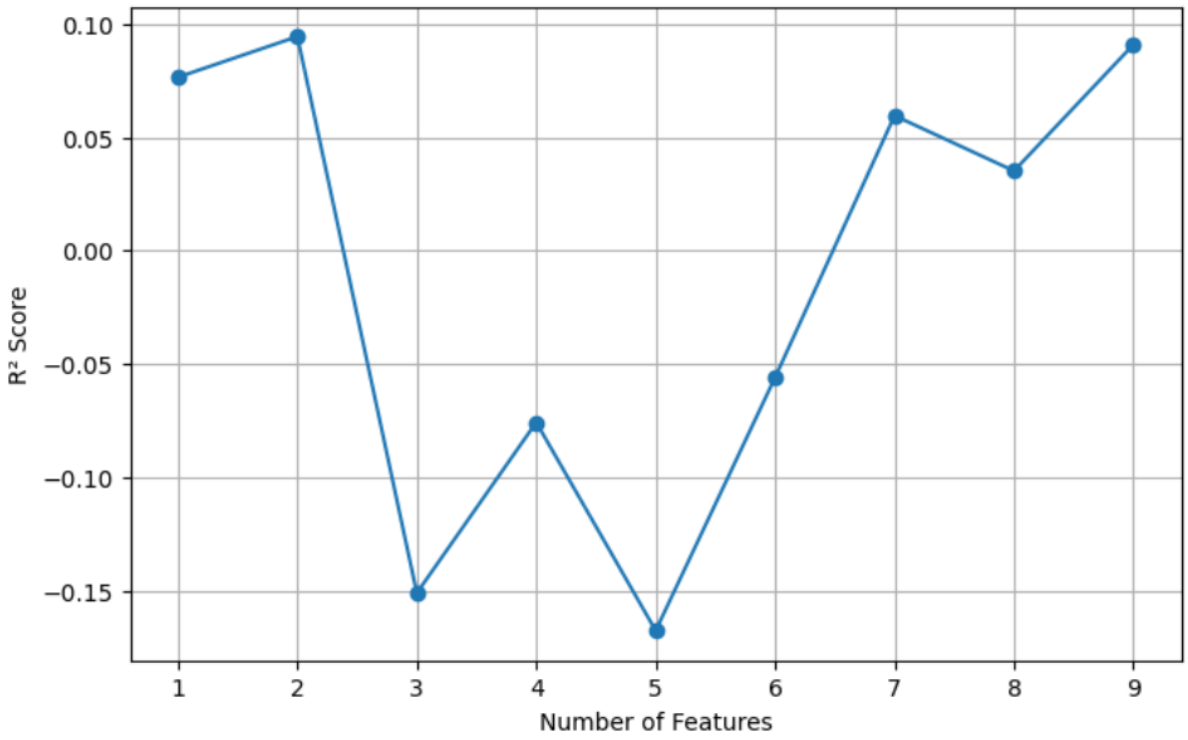

Figure 1: R^2 Score vs. Number of Features Selected

Predicted vs. Actual Prices: Comparison of Naive and Automated Models In the Predicted vs. Actual Prices (Naive) plot (on the left), we can observe a scatter plot where each point represents the predicted price versus the actual price. The scatter points in this plot are less densely packed, and there is a noticeable spread of points, indicating that the naive model's predictions are not very close to the actual values. The black line represents a perfect prediction scenario (where the predicted price equals the actual price), and we can see that the data points generally deviate from this line, demonstrating the inaccuracies of the naive model.

On the other hand, in the Predicted vs. Actual Prices (Automated) plot (on the right), the scatter points are more densely distributed along the entire axis, particularly along the diagonal (where predicted price equals actual price). This shows that the automated model performs better, with predictions that are closer to the actual values. The slope of the line has decreased slightly from the naive model to the automated model. This suggests that the automated model has improved the prediction accuracy, as the predicted prices are now more aligned with the actual values, reducing the overall error.

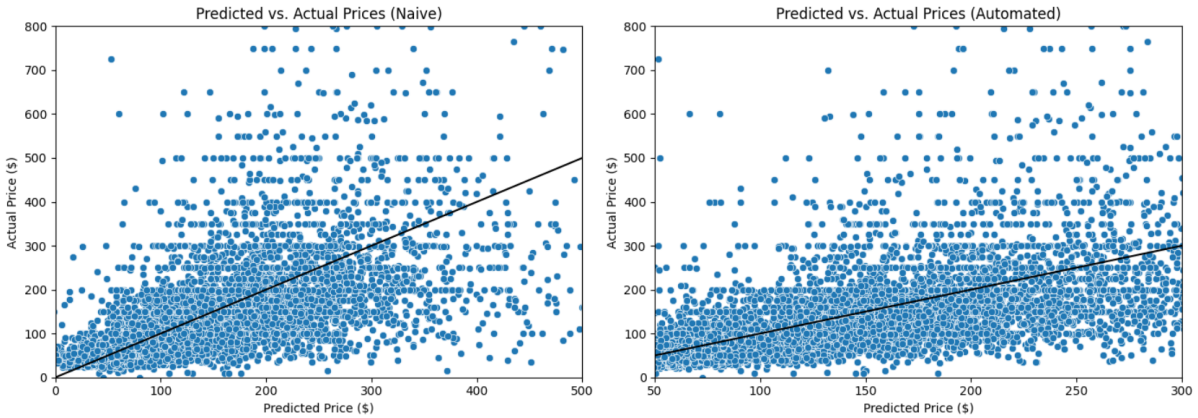

Figure 2: Predicted vs. Actual Prices: Naive vs. Automated Models

4.3 Dataset 2: Bestsellers with Categories

This dataset contains information on bestselling products along with their categories. It is used to predict the sales rank of books based on attributes like price, genre, and author.

Dataset Dimensions:

- **Number of Features:** 7
- **Number of Samples:** 4,000

Target Variable:

- **Sales Rank:** The sales rank of the book in the bestseller list.

Results:

Metric	Naive Solution	Automated Solution
MAE	0.1390	0.1581
MSE	0.0503	0.0474
R^2	0.1227	0.1740
Pearson	0.4394	0.4344

Table 3: Bestsellers with Categories Dataset - Feature Selection Results

4.4 Dataset 3: Global Air Pollution

This dataset contains information on global air pollution levels and environmental factors such as particulate matter (PM2.5), air quality index, and location. It is used to predict air pollution levels based on environmental attributes.

Dataset Dimensions:

- **Number of Features:** 12
- **Number of Samples:** 30,000

Target Variable:

- **PM2.5 (Particulate Matter):** The concentration of particulate matter in the air.

Results:

Metric	Naive Solution	Automated Solution
MAE	15.01	16.81
MSE	884.82	958.77
R^2	0.7223	0.6991
Pearson	0.8500	0.8362

Table 4: Global Air Pollution Dataset - Model Performance

4.5 Dataset 4: Salary Prediction

This dataset contains information on various job features, such as education, experience, and job category. The goal is to predict the salary of employees based on these attributes.

Dataset Dimensions:

- **Number of Features:** 6

- **Number of Samples:** 10,000

Target Variable:

- **Salary:** The annual salary of the employee.

Results:

Metric	Naive Solution	Automated Solution
MAE	10482.01	11723.06
MSE	263729609.29	264482466.92
R^2	0.8969	0.8966
Pearson	0.9481	0.9492

Table 5: Salary Prediction Dataset - Model Performance

4.6 Dataset 5: Sleep Efficiency

This dataset contains information about a group of test subjects and their sleep patterns.

Dataset Dimensions:

- **Number of Features:** 15
- **Number of Samples:** 500

Target Variable:

- **Awakenings:** Number of times each subject wakes up during the night.

Results:

Metric	Naive Solution	Automated Solution
MAE	0.8778	0.8668
MSE	1.3146	1.1159
R^2	0.2461	0.3600
Pearson	0.5670	0.6317

Table 6: Sleep Efficiency Dataset - Model Performance

5 Conclusion

In conclusion, the automated feature selection pipeline significantly improves model performance by removing irrelevant features, ranking the most important features based on SHAP values, and eliminating less useful features through recursive elimination. The evaluation results show improvements in R^2 scores and a reduction in prediction error, making this approach more effective than using all features. Future work could involve extending this approach to other machine learning models, such as random forests or neural networks, as well as testing it with additional datasets to further assess its generalizability.

References

[1] Guyon, I.,
Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3, 1157-1182.

[2] Lundberg, S. M.,
Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Neural Information Processing Systems (NIPS)*.