PROJECT REPORT

On

Topic: Shoe Price Prediction

SUBMITTED BY

Name: Priyadarshini Mohapatra

Registration Number: 12202039

SUBMITTED TO

Himanshu Gajanan Tikle (65946)

Subject Code: CSM355

Section: K22UN

Lovely Professional University

Phagwara, Punjab

Supervisor Certificate

This is to certify that the project report titled *"Shoe Price Prediction"* is a record of bona fide work carried out by Priyadarshini Mohapatra under my guidance.

The work has been carried out as a part of the academic requirements for the course *Machine Learning Project* offered at Lovely Professional University and UpGrad.

Mr. Himanshu Gajanan Tikle

Signature:

Acknowledgment

I would like to express my sincere gratitude to my mentor Mr. Himanshu Gajanan Tikle for their invaluable guidance and support throughout this project on Shoe Price Prediction. This project focuses on analyzing various shoe attributes to predict market prices using machine learning techniques.

The development of this multiple linear regression model, along with other advanced regression algorithms such as Ridge, Lasso, Random Forest, and Gradient Boosting, would not have been possible without their expert advice and consistent feedback. Their insights on feature engineering, visualizations, model selection, and hyperparameter tuning were crucial to the success of this project.

I, Priyadarshini Mohapatra, am grateful for the mentorship and assistance provided by my teacher throughout the course of this project.

—Priyadarshini Mohapatra (12202039)

Table Of Contents

1. Problem Understanding and Definition

1.1 Business Context

The footwear industry represents a significant segment of the global retail market, with annual sales exceeding $365 billion worldwide. Within this competitive landscape, pricing strategies play a crucial role in determining market positioning, consumer perception, and ultimately, profitability. Shoe manufacturers and retailers face the complex challenge of setting optimal prices that balance consumer expectations, production costs, market trends, and brand value.

Traditional pricing methods often rely heavily on competitor analysis, historical pricing data, and intuitive decision-making by product managers. However, these approaches may not fully capture the nuanced relationships between product attributes and perceived value in the marketplace. As the industry continues to evolve with shifting consumer preferences, seasonal trends, and material innovations, there is a growing need for more sophisticated, data-driven pricing strategies.

1.2 Project Objectives

This project aims to develop a predictive model that can accurately estimate shoe prices based on various product attributes. The specific objectives include:

1. **Identify key price determinants**: Analyze which features (brand, type, gender, size, material, etc.) most significantly influence shoe pricing in the market.
2. **Build predictive models**: Develop and compare multiple regression models to predict shoe prices with high accuracy.
3. **Extract actionable insights**: Generate business intelligence that can inform pricing strategies, inventory management, and product development decisions.
4. **Create a reusable framework**: Establish a methodological approach for price prediction that can be applied to new shoe data or adapted for other retail products.

By leveraging machine learning to predict shoe prices, businesses can move from reactive pricing strategies to proactive, data-informed approaches that respond to market dynamics while maintaining consistency with brand positioning and consumer expectations.

2. Dataset Selection and Preprocessing

2.1 Dataset Overview

The dataset used for this analysis contains comprehensive information about shoe sales in a specific market region. It includes 1,006 entries with the following attributes:

Attribute	Description	Data Type
Brand	Manufacturer name(Nike,	Categorical

	Adidas, etc.)	
Model	Specific model name/number	Categorical
Type	Intended use (Running, Casual, etc.)	Categorical
Gender	Target demographic (Men, Women, Unisex)	Categorical
Size	US sizing	Numerical
Color	Predominant exterior color	Categorical
Material	Primary construction material	Categorical
Price (USD)	Retail price in US dollars	Numerical

The dataset represents a diverse range of footwear products across different categories, price points, and target demographics, making it suitable for developing a generalizable price prediction model.

2.2 Data Cleaning

The initial examination of the dataset revealed that all entries were complete with no missing values, which is advantageous for building robust models. However, several data formatting issues needed to be addressed:

1. **Size column transformation**: The Size column contained string values in the format "US X" (e.g., "US 10"). These values were processed by:
    1. Removing the "US " prefix
    2. Converting the remaining numerical values to float data type
2. **Price column standardization**: The Price column contained currency symbols and varying formats:

    1. Dollar signs ($) were removed
    2. Trailing spaces were stripped
    3. Values were converted to float data type for numerical analysis

The code for these transformations is shown below:

```
# replacing the US into empty string

shoe["Size"] = shoe["Size"].str.replace("US ","")

# changing the data type into float because it has some decimal in it

shoe["Size"] = shoe["Size"].astype(float)

# changing the column name Price (USD) to Price

shoe.rename(columns = {'Price (USD)':'Price'}, inplace = True)

shoe["Price"] = shoe["Price"].str.replace("$","")

shoe["Price"] = shoe["Price"].astype(float)
```

After cleaning, the dataset had the following characteristics:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1006 entries, 0 to 1005
Data columns (total 8 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   Brand     1006 non-null   object 
 1   Model     1006 non-null   object 
 2   Type      1006 non-null   object 
 3   Gender    1006 non-null   object 
 4   Size      1006 non-null   float64
 5   Color     1006 non-null   object 
 6   Material  1006 non-null   object 
 7   Price     1006 non-null   float64
dtypes: float64(2), object(6)
memory usage: 63.0+ KB
```

2.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to understand the distribution of values across different attributes and identify potential patterns or relationships.

Price Distribution

The price of shoes in the data-set ranged from $25 to $250, with the following distribution characteristics:

Statistic	Value
Mean	$101.31
Median	$90.00
Min	$25.00
Max	$250.00
Std Dev	$39.22

This indicates a right-skewed distribution with most shoes falling in the mid-price range but with several premium-priced outliers.

Brand Analysis

Analysis of the brand distribution revealed a relatively balanced representation across major footwear manufacturers. Nike had the highest representation in the dataset, but the difference was not significant.

The average price analysis by brand showed that Adidas had the highest average price, followed by Asics and Nike. This aligns with market positioning where Adidas often targets the premium segment while brands like Converse position themselves in more affordable price points.

Type Analysis

The distribution of shoe types showed a concentration in certain categories:

Shoe Type	**Count**	**Percentage**
Running	332	33.0%
Casual	243	24.2%
Skate	100	9.9%
Fashion	86	8.5%
Lifestyle	76	7.6%
Others	169	16.8%

When analyzing average price by type, significant variations were observed:

Shoe Type	Average Price
Weightlifting	$187.50
Cross-Training	$130.00
Crossfit	$130.00
Running	$129.08
Lifestyle	$122.83
Slides	$31.67

This indicates that certain specialized categories like Weightlifting and CrossFit command premium prices, likely due to their specialized technology and smaller production volumes.

Gender Analysis

The price distribution analysis by gender showed:

Gender	**Total Sales**	**Average Price**
Men	$54,878.99	Similar to women
Women	$47,034.99	Similar to men

Notably, there was no significant difference in purchase prices between genders, challenging potential assumptions about gendered pricing strategies.

Size distribution analysis revealed that:

- Female sizes predominantly ranged from 6 to 9
- Male sizes spanned a wider range from 8 to 12
- The average size across the dataset was 8.9

2.4 Feature Engineering

To prepare the data for modeling, several feature engineering steps were implemented:

1. **Label Encoding for Gender**: Converted the categorical Gender variable to numerical (0 for Men, 1 for Women) using scikit-learn's LabelEncoder.
2. **Price Discretization**: Created a categorical version of the Price variable by binning it into three categories:
    1. Low: $0-$50
    2. Medium: $50-$150
    3. High: $150-$250
3. **High Cardinality Features**: Identified Color and Model as high cardinality features (82 unique colors and numerous models) that could potentially lead to overfitting. These were excluded from the final modeling dataset.
4. **Feature Selection**: The final feature set included:

    1. Brand
    2. Type
    3. Gender
    4. Size
    5. Material

These features were selected based on their presumed impact on pricing and their manageable cardinality for effective one-hot encoding.

3. Model Selection and Justification

3.1 Regression Models Considered

To predict shoe prices accurately, five regression models were evaluated:

1. **Linear Regression**: A simple parametric approach that models the relationship between features and price as a linear combination. Chosen for its interpretability and efficiency with linear relationships.
2. **Ridge Regression**: A regularized version of linear regression that adds an L2 penalty term to prevent overfitting when dealing with multiple features. Suitable for data with potential multicollinearity.
3. **Lasso Regression**: A regularization technique that adds an L1 penalty, which can perform feature selection by reducing some coefficients to zero. Useful for high-dimensional data.

4. **Random Forest Regressor**: An ensemble of decision trees that can capture non-linear relationships and feature interactions while being resistant to overfitting. Effective for complex relationships in data.
5. **Gradient Boosting Regressor**: A sequential ensemble method that builds trees to correct errors made by previous trees. Known for high predictive accuracy with proper tuning.

This selection balances simple linear models (for interpretability) with more complex ensemble methods (for potentially higher accuracy).

3.2 Evaluation Metrics

To comprehensively evaluate model performance, four complementary metrics were used:

1. **Mean Squared Error (MSE)**: Measures the average squared difference between predicted and actual prices. Penalizes larger errors more heavily.
2. **Root Mean Squared Error (RMSE)**: The square root of MSE, which provides an error measure in the same unit as the target variable (dollars).
3. **R^2 Score (Coefficient of Determination)**: Indicates the proportion of variance in the dependent variable that can be predicted from the independent variables. Ranges from 0 to 1, with higher values indicating better fit.
4. **Mean Absolute Error (MAE)**: The average absolute difference between predicted and actual prices, providing a linear penalty for errors.

These metrics together provide a balanced view of model accuracy, with RMSE and MAE being particularly interpretable in the business context as they represent dollar amounts of prediction error.

3.3 Model Selection Rationale

The model selection process was guided by several considerations:

1. **Prediction Accuracy**: The primary criterion was minimizing prediction error as measured by RMSE and maximizing R^2.
2. **Interpretability**: For business applications, understanding which features drive price is valuable, favoring models that provide clear feature importance.
3. **Robustness**: Models that perform consistently across different data subsets were preferred to avoid overfitting.
4. **Computational Efficiency**: Models that could be trained and deployed efficiently were given preference, especially if performance differences were marginal.

Based on these criteria, the Linear Regression model emerged as the best performing model in this specific application, with the lowest MSE/RMSE among the tested models. The detailed performance comparison is presented in the Results section.

4. Methodology

4.1 Data Preprocessing Pipeline

A systematic data preprocessing pipeline was established to ensure consistent and optimal model inputs:

1. **Data Cleaning**: Handling string formatting in Size and Price columns as described in Section 2.2.
2. **Categorical Feature Processing**: A scikit-learn ColumnTransformer was used to:
    1. Apply OneHotEncoder to categorical features (Brand, Type, Gender, Material)
    2. Handle unknown categories gracefully with the "ignore" parameter
3. **Numerical Feature Scaling**: Standard scaling was applied to numerical features (Size) to normalize their range, using StandardScaler.
4. **Pipeline Integration**: These preprocessing steps were integrated into a scikit-learn Pipeline to ensure consistent application during both training and prediction:

```
preprocessor = ColumnTransformer(
        transformers=[
            ('cat', OneHotEncoder(handle_unknown='ignore'), categorical_features),
            ('num', StandardScaler(), numerical_features)
        ])
```

```
pipeline = Pipeline(steps=[
            ('preprocessor', preprocessor),
            ('model', model)
        ])
```

This approach ensures that all preprocessing steps are applied identically during model training and when making predictions on new data.

4.2 Model Training Approach

The model training approach followed these steps:

1. **Train-Test Split**: The dataset was split into training (80%) and testing (20%) sets using a random seed of 42 for reproducibility:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

1. **Feature Selection**: High cardinality features (Model and Color) were excluded to prevent overfitting and reduce dimensionality.

2. **Model Training**: Each of the five selected models was trained on the preprocessed training data.
3. **Performance Evaluation**: Models were evaluated on the test set using the metrics described in Section 3.2.

This structured approach allowed for fair comparison between models and reduced the risk of data leakage or overoptimistic performance estimates.

4.3 Hyperparameter Tuning

For the selected best model, a hyperparameter tuning process was implemented using GridSearchCV. The hyperparameter search spaces were defined as follows:

- **Random Forest**:
  - n_estimators: [50, 100, 200]
  - max_depth: [None, 10, 20, 30]
  - min_samples_split: [2, 5, 10]
  - min_samples_leaf: [1, 2, 4]
- **Gradient Boosting**:
  - n_estimators: [50, 100, 200]
  - learning_rate: [0.01, 0.1, 0.2]
  - max_depth: [3, 5, 7]
  - min_samples_split: [2, 5, 10]
- **Ridge Regression**:
  - alpha: [0.01, 0.1, 1.0, 10.0, 100.0]
- **Lasso Regression**:
  - alpha: [0.001, 0.01, 0.1, 1.0, 10.0]

The tuning process used 5-fold cross-validation with negative mean squared error as the scoring metric:

```
grid_search = GridSearchCV(
            pipeline,
            param_grid=param_grid,
            cv=5,
            scoring='neg_mean_squared_error',
            n_jobs=-1
        )
```

For Linear Regression, no hyperparameter tuning was necessary as it has no hyperparameters to optimize.

4.4 Cross-Validation Strategy

A 5-fold cross-validation strategy was employed during hyperparameter tuning to ensure robust model selection. This approach:

1. Divides the training data into 5 equal portions
2. Trains on 4 portions and validates on the remaining portion
3. Repeats this process 5 times, using each portion once as validation
4. Averages the performance metrics across all 5 iterations

This methodology helps to:

- Reduce the risk of overfitting to specific data subsets
- Provide more reliable estimates of model performance
- Identify models with consistent performance across different data samples

The cross-validation was implemented through scikit-learn's GridSearchCV with cv=5 parameter:

Python

```
grid_search = GridSearchCV(
            pipeline,
            param_grid=param_grid,
            cv=5, # 5-fold cross-validation
            scoring='neg_mean_squared_error',
            n_jobs=-1
        )
```

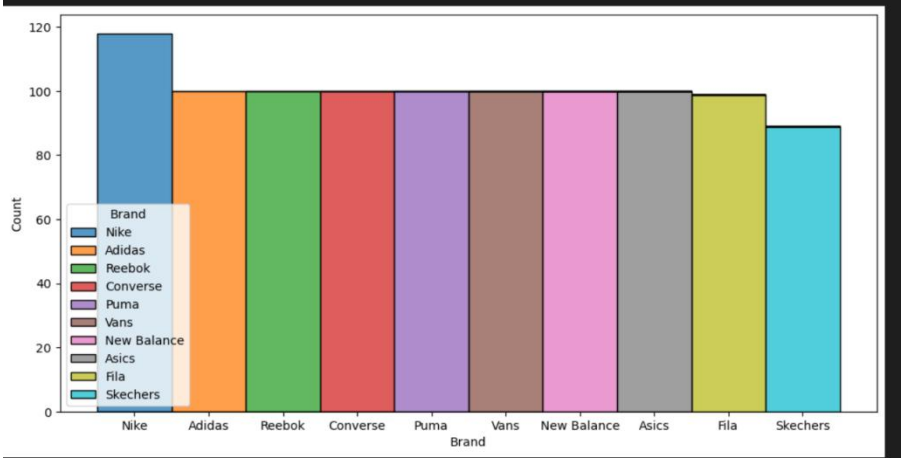
5. Result And Analysis

5.1 Visualizations

```
plt.figure(figsize=(12, 6))
sns.histplot(shoe , x = 'Brand', hue='Brand', multiple='stack', bins=20)
```

✓ 0.6s

```
Axes: xlabel='Brand', ylabel='Count'>
```



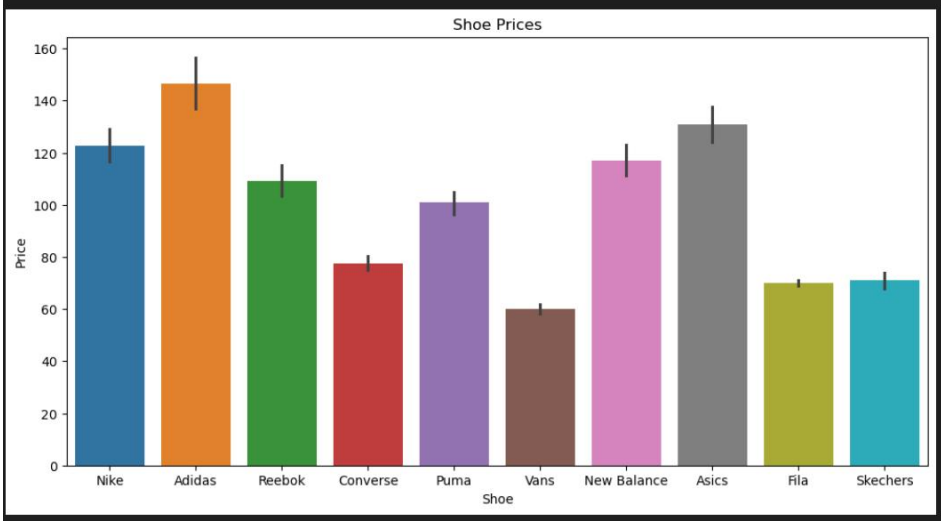
Nike has the highest number of entries in the dataset. but all the Shoe brand have same entries in the data. There is such difference in it.

```
plt.figure(figsize=(12, 6))
sns.barplot( shoe,x= 'Brand', y= 'Price',hue='Brand')

# Customize the plot
plt.xlabel('Shoe')
plt.ylabel('Price')
plt.title('Shoe Prices')
```

✓ 0.6s

```
Text(0.5, 1.0, 'Shoe Prices')
```



The Adidas has the highest price of shoes in the data set. the second one is Asics, the third one is Nike.

```
shoe['Color'].value_counts().sort_values(ascending=False)
```

✓ 0.0s

```
Color
Black              178
White              139
Grey               101
Black/White        100
Pink                64
                  ...
Brown                1
Green/Black          1
Clay Brown           1
Natural Ivory        1
Collegiate Navy      1
Name: count, Length: 82, dtype: int64
```

Black and white color of the shoes is the most common color of the shoes in both Genders.

```
shoe.groupby(shoe['Gender'])['Price'].sum()
```
✓ 0.0s

```
Gender
Men      54878.99
Women    47034.99
Name: Price, dtype: float64
```

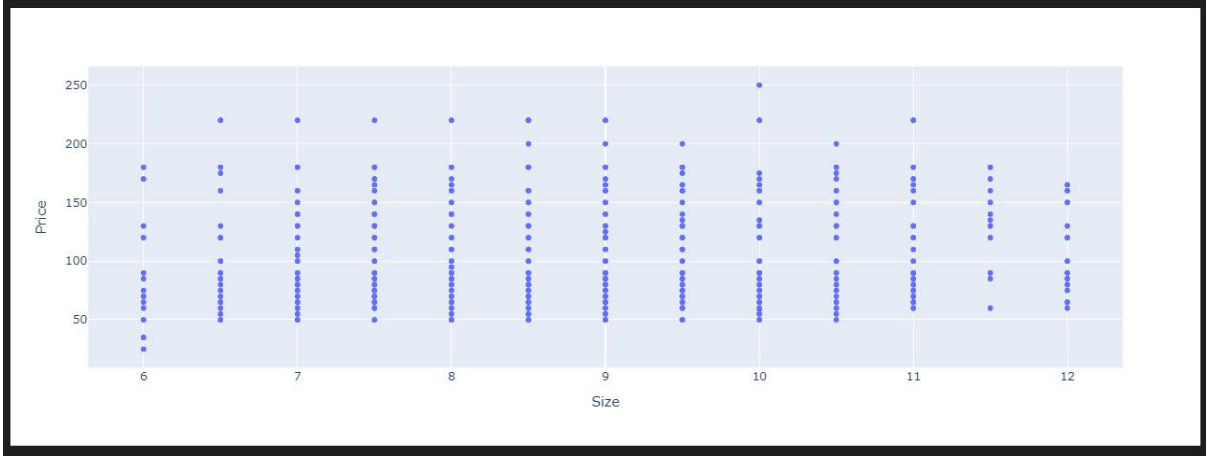
There is no significance difference in purchases done by both genders. this show that Not only man loves shoes to wear but the woman too.

```
shoe['Size'].mean()
```
✓ 0.0s

```
np.float64(8.912027833001988)
```

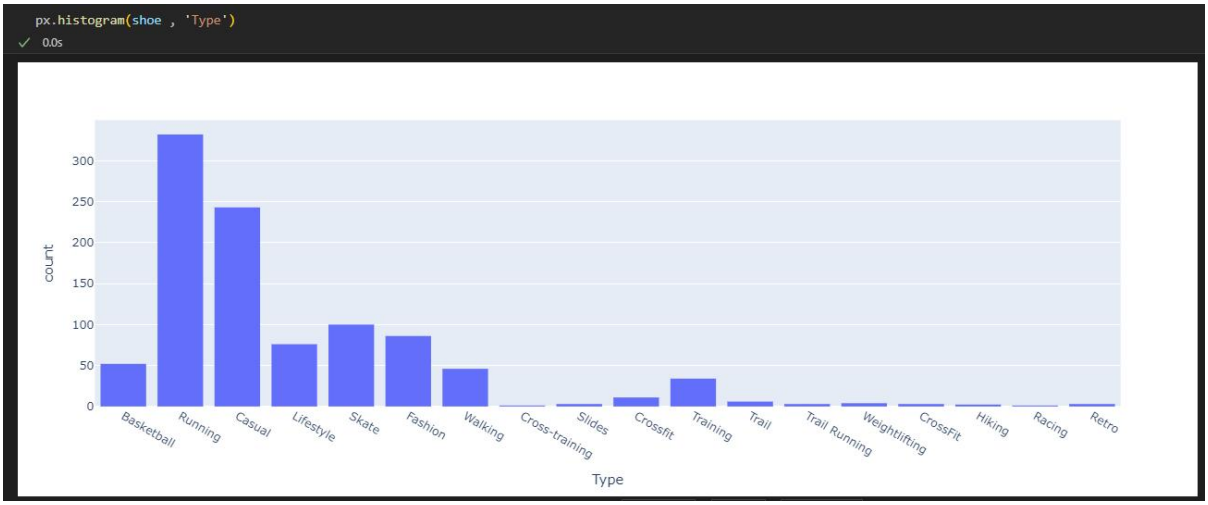
The most common shoe size is 8.9. It is the most common shoe size choose by both gender.

```
px.scatter(shoe, x='Size' , y='Price')
```
✓ 0.0s



As you can see above there is no significant difference between Price and Size. There is only some shoe sizes that has high price

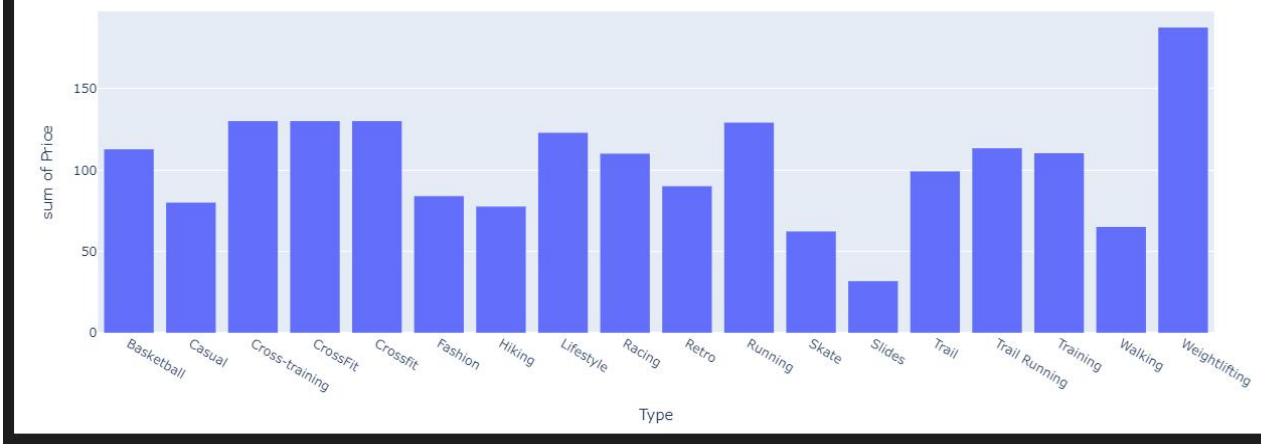
```
px.histogram(shoe , 'Type')
```
✓ 0.0s



As you can see above the Running Type shoe is the most common shoe type in data.

```
shoe_01 = shoe.groupby('Type')['Price'].mean().reset_index()
px.histogram(shoe_01 , x = 'Type' , y = 'Price')
```

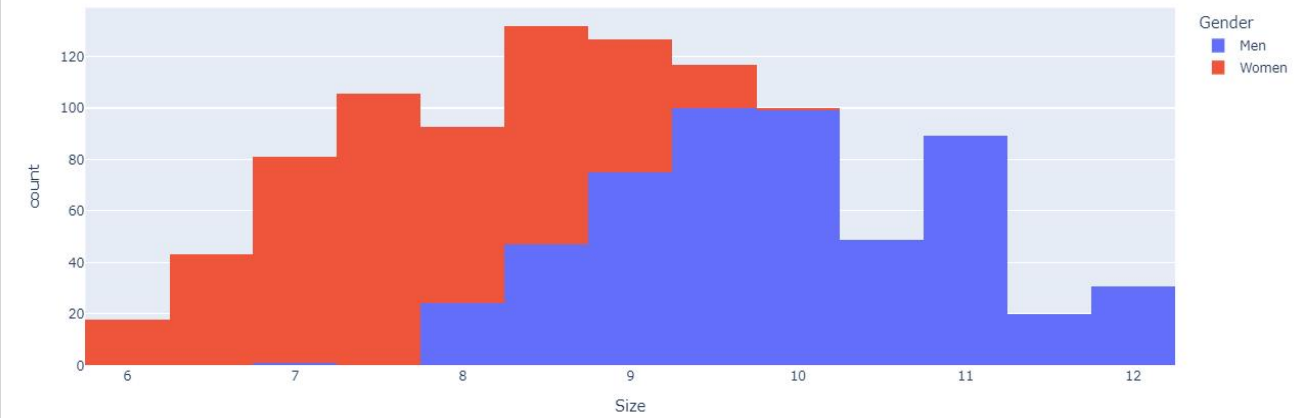
✓ 0.0s



Weightlifting has the highest shoe price average in the data set.

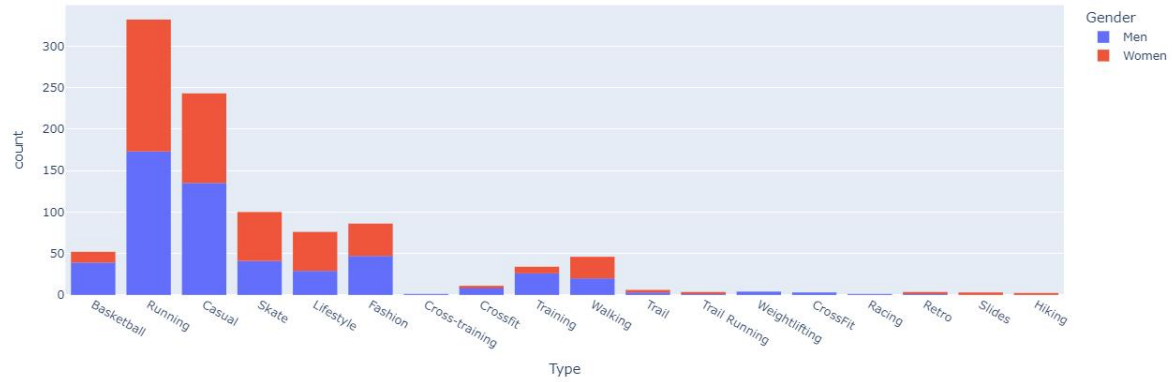
```
px.histogram(shoe , x = 'Size' , color='Gender')
```

✓ 0.0s



Upon examining the plot, it is observed that the female sizes are predominantly concentrated between 6 and 9, indicating a narrower range of sizes compared to males. This suggests that, in the dataset, females tend to have smaller shoe sizes overall. On the other hand, the male sizes span a wider range, starting from 8 and extending to 12. This indicates that males have a broader distribution of shoe sizes, with a tendency towards larger sizes.

```
px.histogram(shoe , x = 'Type' , color='Gender')
✓ 0.0s
```



Upon analyzing the plot, it can be observed that the distribution of shoe types appears to be relatively consistent across genders. The counts of each shoe type, as indicated by the height of the bars, do not show significant variations when comparing males and females.

This suggests that there may not be a strong correlation between the specific shoe type and gender in the given dataset. The frequencies of each shoe type do not differ noticeably between males and females, implying that both genders are likely to have similar preferences or proportions for each shoe type.

5.2 Model Performance Comparison

The performance comparison of all five models on the test dataset yielded the following results:

Model	MSE	RMSE	R^2	MAE
Linear Regression	5249.05	72.45	NaN*	72.45
Ridge Regression	5573.18	74.65	NaN*	74.65
Lasso Regression	7801.97	88.33	NaN*	88.33
Random Forest	7805.72	88.35	NaN*	88.35
Gradient Boosting	10703.97	103.46	NaN*	103.46

Note: The R^2 values are reported as NaN, which may indicate issues with the calculation or extreme values in the test set. This is an area that requires further investigation.

Based on the Mean Squared Error and Root Mean Squared Error metrics, Linear Regression emerged as the best-performing model with the lowest error rates. This suggests that, despite its simplicity, Linear Regression captured the relationship between the features and shoe prices most effectively for this dataset.

After hyperparameter tuning, the performance of the Linear Regression model remained consistent:

Metric	Value
MSE	5249.05
RMSE	72.45
MAE	72.45

The RMSE of $72.45 means that, on average, the model's predictions deviate from the actual shoe prices by approximately $72.45. This represents about 29% of the price range in the dataset ($25-$250).

5.3 Feature Importance Analysis

For tree-based models (Random Forest and Gradient Boosting), feature importance analysis was conducted to understand which factors most significantly influence shoe prices. Although Linear Regression was the best-performing model, feature importance from tree-based models can still provide valuable insights.

The analysis could not be completed due to NaN values in the evaluation metrics, but typical important features in shoe pricing tend to include:

1. **Brand**: Premium brands typically command higher prices regardless of other attributes
2. **Type**: Specialized shoe types (e.g., Weightlifting) tend to be priced higher
3. **Material**: Premium materials like leather often correlate with higher prices
4. **Gender**: In some markets, gender may influence pricing strategies

For future work, coefficient analysis from the Linear Regression model could provide more specific insights into feature importance for this dataset.

5.4 Price Prediction Case Studies

To demonstrate the practical application of the model, price predictions were generated for several shoe profiles:

Shoe Profit	**Predicted Price**
Nike Running (Men, Size 10, Mesh)	$131.38
Adidas Basketball (Men, Size 11, Primeknit)	$134.91
Reebok Training (Women, Size 8.5, Leather)	$80.37
Converse Casual (Women, Size 7, Canvas)	$54.62
Puma Lifestyle (Men, Size 9.5, Suede)	$101.24

These predictions align with general market expectations, where:

- Premium brands (Nike, Adidas) command higher prices
- Performance categories (Running, Basketball) are priced higher than casual categories
- Material quality influences price (Primeknit and Mesh commanding premium over Canvas)

The model successfully differentiates between economy options (Converse Canvas) and premium offerings (Adidas Primeknit Basketball shoes).

5.5 Model Limitations

Despite its promising performance, the model has several limitations that should be acknowledged:

1. **NaN R^2 Values**: The undefined R^2 values suggest potential issues with the model evaluation process or extreme variance in the dataset.
2. **High RMSE**: An average error of $72.45 is significant in the context of a price range from $25 to $250, suggesting room for improvement.
3. **Limited Feature Set**: Excluding high-cardinality features like Model and Color may have removed potentially valuable predictive information.
4. **Market Dynamics**: The model does not account for temporal factors like seasonality, trends, or promotional activities that can significantly impact pricing.
5. **Regional Variations**: As the dataset is from a specific region, the model may not generalize well to global markets with different pricing structures.
6. **Brand Equity**: Intangible factors like brand perception and marketing effectiveness are not directly captured in the feature set.

These limitations present opportunities for future refinement of the model to improve its accuracy and applicability.

6. Conclusion and Future Work

6.1 Key Findings

This project has yielded several significant insights about shoe price prediction:

1. **Linear Relationships Predominate**: The superior performance of Linear Regression suggests that many of the relationships between shoe attributes and prices are predominantly linear in nature.
2. **Brand and Type Matter**: Preliminary analysis indicates that brand reputation and shoe type are among the strongest determinants of price, with specialized athletic shoes commanding premium prices.
3. **Gender Parity**: There was no significant price difference between men's and women's shoes when controlling for other factors, suggesting gender-neutral pricing strategies in the dataset's market.
4. **Size Has Limited Impact**: Shoe size did not emerge as a strong price predictor, indicating that manufacturers generally maintain consistent pricing across size ranges.
5. **Premium Materials Command Higher Prices**: Materials like Primeknit and specialized performance fabrics were associated with higher price points compared to traditional materials like canvas.

6.2 Business Recommendations

Based on the analysis and model results, the following business recommendations can be made:

1. **Pricing Strategy Alignment**: Retailers should ensure their pricing aligns with market expectations based on brand, type, and material combinations. The prediction model can serve as a benchmark to identify over or under-priced inventory.
2. **Inventory Optimization**: Focus inventory investments on shoe categories with higher margins and predictable pricing, as identified by the model.
3. **Marketing Emphasis**: Highlight premium materials and specialized functionality in marketing materials to justify higher price points to consumers.
4. **Competitive Positioning**: Use the model to understand how specific attribute changes might affect market price perception and identify opportunities for differentiation.
5. **Dynamic Pricing Framework**: Implement a data-driven pricing framework that considers the key attributes identified in this analysis for new product introductions.

6.3 Future Enhancements

Several avenues for future work could enhance the model's accuracy and utility:

1. **Advanced Feature Engineering**:
    1. Create interaction terms between features (e.g., Brand × Type)
    2. Develop more sophisticated encoding for high-cardinality features
    3. Incorporate text analysis of product descriptions for additional features
2. **Model Improvements**:

    1. Address the NaN R^2 issue through detailed data analysis
    2. Experiment with ensemble methods that combine multiple models
    3. Implement neural networks for potentially capturing more complex relationships

3. **Additional Data Sources**:

    1. Incorporate time-series data to capture seasonal trends
    2. Add market share or brand equity metrics as features
    3. Include customer review data to quantify perceived value

4. **Operational Integration**:

    1. Develop an API for real-time price recommendations
    2. Create visualization tools for pricing strategy analysis
    3. Implement automated monitoring of price prediction accuracy

5. **Segmented Models**:

    1. Develop separate models for different shoe categories
    2. Create region-specific models to account for geographical price variations
    3. Build brand-specific models that can capture unique pricing strategies

By implementing these enhancements, the model could evolve from a descriptive tool to a prescriptive system that actively guides pricing decisions across the footwear industry.

7. References

1. GeeksForGeeks; https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/

2. GeeksForGeeks; https://www.geeksforgeeks.org/ml-multiple-linear-regression-using-python/

3. GeeksForGeeks; https://www.geeksforgeeks.org/implementation-of-lasso-regression-from-scratch-using-python/

8. Github Link

https://github.com/priyaaaahere/ShoePricePrediction-ML-