Introduction

This research aims to create a predictive model to forecast the prices of silver that is statistically accurate by using the historical predictive data from 2016 to 2026. This research aims to find important predictors for the model by checking the assumption of a linear regression model to improve its accuracy by applying necessary correcting steps to the models using a pipeline from a machine learning approach that combines classical statistical testing with regression regularization for a better result.

Description and Data Preprocessing

The data contains time-series forecasts of silver price predictions, with columns labeled as Predicted Price, Lower Bound, Upper Bound, and Date. The Date column has been converted to numerical time features (Year and Month) because time-series data can have patterns related to the months of the year. The target variable to be used in modeling is Predicted Price, whereas Lower Bound, Upper Bound, Year, and Month would be used as predictors.

Before carrying out modeling, checks for missing and inconsistent data were also conducted on the data set. Imputation for numerical variables using the median was then undertaken in this research to withstand the effect of outliers in the data set. Another consideration in modeling was making the model less sensitive using scale transformation and power transform Fucking.

Choice of Regression Technique

Multiple Linear Regression (MLR) is dually important as it is originally used for studying the relationship between predictors and the variable of interest, as well as performing statistical inference on that relationship. With an MLR model, estimates of the regression coefficients and their corresponding p-values can be determined, which is particularly useful for identifying which predictors are statistically significant and important for testing theories.

However, after performing a diagnostic test, multicollinearity was observed among variables Lower Bound and Upper Bound, which is common since they are all obtained from price bounds. This can be remedied by using Ridge Regression as a final prediction algorithm since Ridge Regression implements an L2 norm that tries to reduce variance by limiting large coefficients while retaining features. This is necessary because of the nature of stock variables to some degree.

Identification of Significant Predictors

The data was tested for statistical significance by utilizing the results of the Ordinary Least Squares (OLS) regression technique. The results revealed the statistical significance of the factors **Lower Bound** and **Upper Bound**, which act as direct indicators for the value of the expected silver price. The statistical significance of the **Year** factor was established as a result of the increment in the silver prices over the years. The statistical significance of the **Month** factor was demonstrated to be lower in importance but still relevant.

These results are consistent with financial reasoning in that silver prices are very sensitive to forecast bounds rather than to month-level dynamics.

Checking Assumptions of Multiple Linear Regression

There was a thorough diagnostic analysis that helped assess the classical assumptions of the multiple linear regression model.

The linearity was assessed by looking at the behavior of residuals and model fit, and even if a linear relationship was largely present, small deviations did occur. The normality test of residuals showed slight skewness, and hence, the error terms did not follow a perfect normal distribution. The homoscedasticity assumption was formally assessed using a Breusch-Pagan test and showed heteroscedasticity in error variance, implying that variance is not consistent across predicted values. Multicollinearity was measured using Variance Inflation Factor (VIF), and large VIF values confirmed high correlation among predictors. The independence of error terms was fairly fulfilled after considering variables related to time.

Violations of such nature are prevalent in financial as well as commodity prices, and if not handled, they tend to have a negative effect on inference as well as prediction accuracy, respectively.

Assumption	Test Used	Result	Remedy
Linearity	Residual diagnostics	Minor deviation	Power transformation
Normality	Residual distribution	Mild skewness	Yeo-Johnson transforms
Homoscedasticity	Breusch–Pagan test	Violation detected	Transformation + Ridge
Multicollinearity	VIF	High correlation	Ridge regularization
Independence	Time-based structure	Acceptable	Time variables added

Remedies and Pipeline Implementation

For the removal of identified violations, a machine learning process was developed and employed. Power transformation via the Yeo-Johnson procedure was employed for the removal of skewness and stabilize variance inflation for improved normality and homoscedasticity. Standardization was employed for ensuring that each predictor was contributing equally to the model, which made the penalty term of the Ridge regression work properly.

Ridge regression was also incorporated in the pipeline to reduce the multicollinearity problem among the coefficients of the variables by forcing them to be close to each other. The whole data processing and model development stage was implemented in a pipeline structure to ensure all data transformations were performed in the similar manner during the training and testing phases. The pipeline structure also avoids the leakage of data and improves the reproducibility of the work, which is a major need in today's predictive analysis.

Model Performance and Evaluation

The Ridge regression model performed well in terms of making predictions, as its coefficient of determination, R^2, was high, and the root mean squared error, RMSE, was low on the test set. Specifically, the generalization capability of the model was improved, and the values of the coefficients were stable compared to the unregularized multiple linear regression model. The presence of the preprocessing modules improved the reliability of the models.