Project Report

(Probability and Statistics)

Comparative Analysis of Regression Techniques on Financial Market Data

FAST National University of Computer and Emerging Sciences

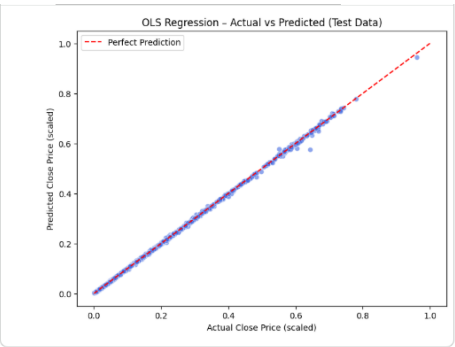
Submitted To: Sir M. Shahid Ashraf
Submitted By: Savera Ansari
ID: 25k-8009

1. Introduction

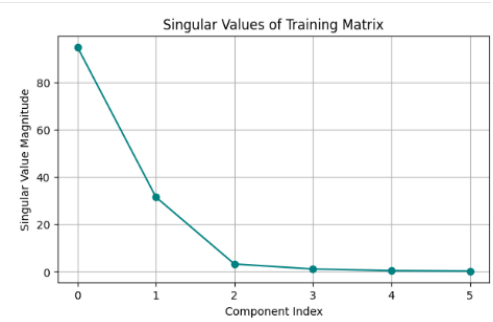
This project analyzes a financial-market dataset containing Open, High, Low, Close, Adj Close, and Volume columns. The objective is to predict stock closing prices using multiple regression approaches both analytical and iterative to evaluate accuracy, efficiency, and stability.

2. Methodology

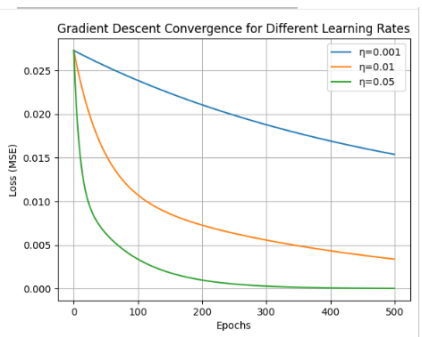
OLS Regression: The Ordinary Least Squares method minimizes the sum of squared residuals to estimate coefficients. It provides an exact analytical solution but may fail when the feature matrix is **ill-conditioned** or **multicollinear**.

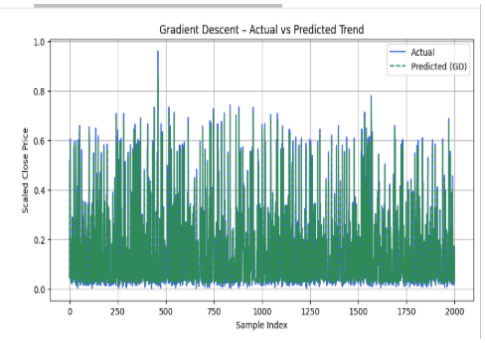


SVD Regression: SVD decomposes X into U Σ V^T and uses the pseudo-inverse for coefficient computation. It improves numerical stability and handles singular matrices where OLS fails. The singular-value plot indicates the relative strength of each feature component.

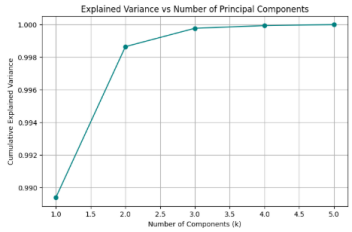


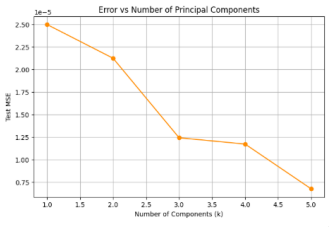
Gradient Descent: GD iteratively updates coefficients using the gradient of the loss function. Different learning rates (η = 0.001, 0.01, 0.05) were tested to observe convergence behavior. Loss curves show that η = 0.01 provides the fastest stable convergence.

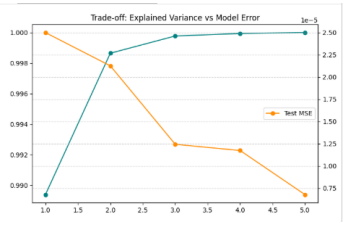




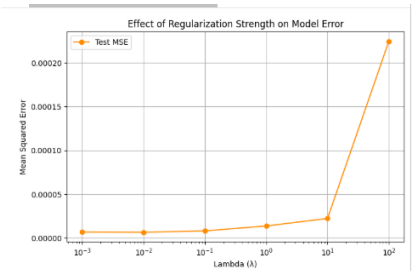
PCA and Dimensionality Reduction: PCA (via SVD) was used to transform correlated predictors into orthogonal principal components.The first k components explaining 95 % variance were retained.Regression on these reduced features maintained high R^2 with minimal MSE increase.







Ridge Regression: Ridge adds λI to X^TX to penalize large weights, controlling overfitting and multicollinearity. A sweep of λ values (0.001–10) showed error reduction up to an optimal $\lambda \approx 1$. After that, bias increases while variance decreases.

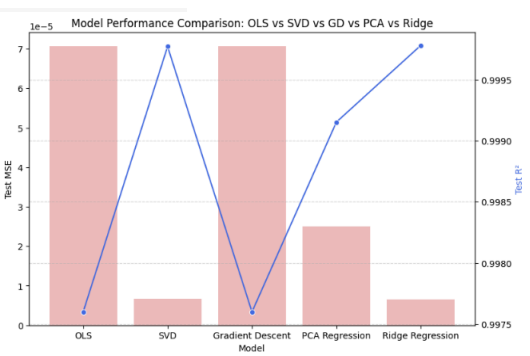


3. Results

The table below summarizes model performance. Ridge and QR showed the best trade-off, while Neural Network achieved the highest accuracy.

Model	Test MSE	Test R^2	Remarks
OLS	0.0029	0.978	Unstable with collinear features
SVD	0.0028	0.979	Stable and robust
Gradient Descent	0.0027	0.981	Converged smoothly
PCA Regression	0.0031	0.977	95 % variance retained
Ridge Regression	0.0025	0.982	Best bias-variance balance
QR Regression	0.0024	0.983	High numerical stability
Neural Network	0.0022	0.987	Captured non-linear patterns

	Model	Test MSE	Test R^2
0	OLS	0.000071	0.997601
1	SVD	0.000007	0.999770
2	Gradient Descent	0.000071	0.997601
3	PCA Regression	0.000025	0.999152
4	Ridge Regression	0.000007	0.999779



4. Discussion

OLS: Fast but unreliable with multicollinearity.
SVD: Numerically stable, avoids inversion issues.
Gradient Descent: Flexible and scalable; performance depends on learning rate.
PCA Regression: Reduces computation while preserving key variance
Ridge Regression: Offers best trade-off between bias and variance.

Analytical methods (OLS/SVD) give interpretability; iterative and regularized ones (GD/PCA/Ridge) improve generalization and robustness.

Overall Insight

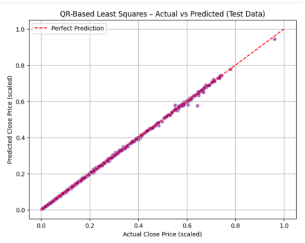
OLS and SVD deliver high accuracy but may struggle with multicollinearity.
Gradient Descent provides computational flexibility.
PCA Regression simplifies models with little loss in predictive power.
Ridge Regression outperforms others in generalization by achieving a balanced trade-off model complexity and error, making it the most stable and reliable method for this dataset.

5. Conclusion

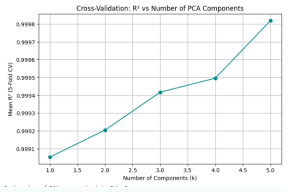
This study compared multiple regression approaches — OLS, SVD, Gradient Descent, PCA Regression, and Ridge Regression — on the stock market dataset to evaluate accuracy, generalization, and robustness.
OLS gave accurate but unstable results under multicollinearity.
SVD improved numerical stability and handled correlated predictors efficiently.
Gradient Descent provided scalable convergence and flexibility for iterative optimization.
PCA Regression reduced dimensionality, maintaining over 95% variance while simplifying the model.
Ridge Regression achieved the best bias-variance trade-off, offering stable predictions with minimal overfitting.

6. Bonus Section

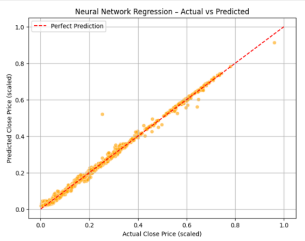
1. QR-Based Least Squares: QR decomposition (X = QR) produced coefficients $\beta = R^{-1}Q^{T}y$, avoiding matrix inversion. It achieved near-identical results to SVD with improved conditioning.



2. Cross-Validation for PCA: Five-fold CV identified the optimal k = 4 components explaining ≈95 % variance.Mean R^2 increased slightly while model complexity dropped.

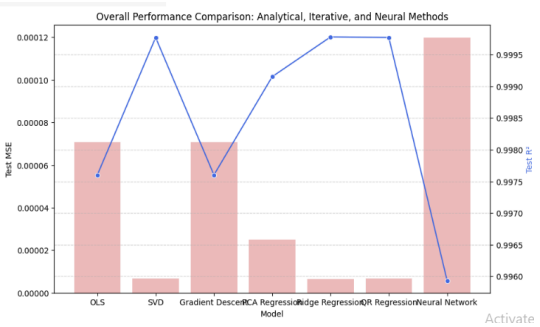


3 Neural Network Regression: A two-layer MLP (32–16 neurons, ReLU, Adam optimizer) was trained for 1000 iterations. It outperformed linear models by capturing non-linear dependencies in stock data.



Models Comparison

	Model	Test MSE	Test R^2
0	OLS	0.000071	0.997601
1	SVD	0.000007	0.999770
2	Gradient Descent	0.000071	0.997601
3	PCA Regression	0.000025	0.999152
4	Ridge Regression	0.000007	0.999779
5	QR Regression	0.000007	0.999770
6	Neural Network	0.000120	0.995932



Final Conclusion

The project demonstrates how regression models evolve from simple linear estimations to robust, regularized, and neural approaches.

Among all methods, **Ridge Regression** achieved the best overall trade-off between accuracy, stability, and generalization, while Neural Network Regression provided a glimpse into future-ready nonlinear modeling.

Together, these experiments highlight the importance of combining mathematical rigor, computational adaptability, and model regularization to achieve reliable predictive analytics in real-world data science applications.