Mobile Price Prediction

Ramakrishna Mission Vivekananda Educational & Research Institute
Belur Math, Howrah, West Bengal
Department of Computer Science
Machine Learning – Course Project Report

Student Name: Arkaprava Jas Student Id: B2530072

---

1 Problem Statement

The goal of this project is to build a machine learning model that predicts the **price range of mobile phones** based on their specifications such as RAM, battery power, screen size, camera quality, and other features. The model helps manufacturers and consumers understand how various attributes influence price segments.

2 Proposed Methodology

1. Objective

The primary objective of the proposed methodology is to develop and evaluate multiple regression-based predictive models to determine the best-performing method for the given dataset based on three performance metrics — **Mean Squared Error (MSE)**, **Mean Absolute Error (MAE)**, and **Coefficient of Determination (R^2)**. The goal is to compare traditional, regularized, and optimization-based learning methods on a common evaluation framework.

2. Workflow Overview

The proposed methodology follows a structured pipeline consisting of the following key stages:

Step 1 — Data Preprocessing

The dataset was imported, cleaned, and prepared for modeling. Feature scaling techniques such as **Min–Max Scaling** and **Standardization** were applied depending on the model (e.g., Gradient Descent methods require normalization for faster convergence).

The dataset was split into:

- Training set: 70%
- Validation set: 20%
- Test set: 10%

This ensures proper evaluation of generalization and avoids overfitting.

Step 2 — Model Development

A set of twelve models were implemented and trained using the training data: **Linear Regression (Ordinary Least Squares),Polynomial Regression (Degree 2, 3, and 4). Batch Gradient Descent (BGD), Stochastic Gradient Descent (SGD), Mini-Batch Gradient Descent (MBGD),Ridge Regularization,Lasso Regularization,Elastic Net Regression,Normal Equation Solution, Singular Value Decomposition (SVD)**

Each model was trained independently using the same training data to ensure fair comparison.

Step 3 — Validation and Testing

After training, each model was evaluated on:

- **Validation set (20%)** — Used to tune hyperparameters and assess model generalization.
- **Test set (10%)** — Used to verify final performance on unseen data.

The following three performance metrics were computed for every model:

- **MSE (Mean Squared Error)** — measures average squared prediction error.
- **MAE (Mean Absolute Error)** — measures average absolute deviation.
- **R^2 (Coefficient of Determination)** — measures model fit quality.

Step 4 — Performance Comparison

A comprehensive performance table was generated summarizing:

- Training, Validation, and Test errors (MSE, MAE, R^2)
- Visualizations using bar charts for all metrics

This allowed direct comparison of model efficiency, bias–variance tradeoff, and generalization ability.

Step 5 — Model Selection and Interpretation

Based on the analysis, Model will be identified as the best-performing model, achieving:

- Lowest MSE and MAE
- Highest R^2
- Stable performance across validation and test datasets

3 Dataset Details

The dataset contains mobile phone specifications like: **Brand, Model ,Storage,Ram,Screen Size (inches),Camera (MP),Battery Capacity(mAh),Price**

Target variable: **price_range** The dataset was taken from the **Kaggle Mobile Price Classification** dataset.
The Shape of the Dataset is (408,7). After Removing the Duplicate data/input the Valid dataset is (381,8). Also at the latent stage Model Name column had been removed

Table 1: Dataset Summary

#	Column Name	Non-Null Count	Data Type
0	Brand	381	object
1	Storage(GB)	381	int64
2	RAM(GB)	381	int64
3	ScreenSize(inches)	381	float64
4	Camera(MP)	381	int64
5	BatteryCapacity(mAh)	381	int64
6	Price($)	381	float64
7	Brand_Encoded	381	int64

4 Comparative Analysis & Results

1. Performance Metrics

The comparison is based on three performance metrics:

- **Mean Squared Error (MSE)** – Measures the average of squared prediction errors. It penalizes larger errors more heavily.
- **Mean Absolute Error (MAE)** – Measures the average magnitude of prediction errors without considering their direction.
- **Coefficient of Determination (R^2)** – Indicates how well the model explains the variance in the target variable; higher values imply better fit.

2. Analysis based on Validation

Validation Results: Validation results reflect each model's ability to generalize to unseen data. The three key metrics used are:

- **Validation MSE** — measures average squared prediction error (lower is better).
- **Validation MAE** — measures average absolute error (lower is better).
- **Validation R^2** — measures the proportion of variance explained by the model (higher is better).

Interpretation

- **Polynomial Regression (Degree 2)** did not outperform the linear models, suggesting that the dataset's relationships are predominantly linear rather than non-linear.
- **Higher-degree polynomials (3 & 4)** performed extremely poorly with very high MSE and negative R^2 values, confirming severe overfitting.
- **Gradient Descent-based models** (Batch, Stochastic, Mini-Batch) converged to the same solution as Linear Regression, validating the correctness and consistency of their implementations.

Table 2: Validation Error Metrics for Different Regression Models

Model	**Validation MSE**	**Validation MAE**	**Validation R^2**
Linear Regression	34624.112	129.294	0.676
Polynomial Regression (Degree 2)	36389.717	127.353	0.659
Polynomial Regression (Degree 3)	193689.156	186.123	-0.814
Polynomial Regression (Degree 4)	97473595.150	1406.605	-911.828
Batch Gradient Descent	31495.716	122.001	0.705
Stochastic Gradient Descent	34640.155	129.109	0.676
Mini-Batch Gradient Descent	33379.416	127.690	0.687
Polynomial Ridge Regression (Degree 2)	34624.079	129.294	0.676
Polynomial Lasso Regression (Degree 2)	34624.108	129.294	0.676
Elastic Net Regression	34624.100	129.294	0.676
Normal Equation	32689.208	123.081	0.694
SVD	32689.208	123.081	0.694

- **Regularized models** (Ridge, Lasso, Elastic Net) neither improved nor degraded performance significantly — implying low multicollinearity and absence of noisy or redundant features in the dataset.
- **Normal Equation and SVD** slightly outperformed other models due to exact analytical computation and better numerical precision, yielding the best validation R^2.

Best-Performing Models best on Validation

Normal Equation and **SVD** achieved the best overall performance on the validation data, with:

$$\text{Validation } R^2 = 0.694, \quad \text{MSE} \approx 32689, \quad \text{MAE} \approx 123$$

These models generalize slightly better than the rest, providing the most effective trade-off between accuracy and stability.

Polynomial Regression beyond Degree 2 exhibited clear signs of overfitting, while Linear and Regularized Regression methods (Ridge, Lasso, Elastic Net) remained robust and consistent across both validation and test datasets.

3.Analysis Best on K(=5)-Fold Validation

K-Fold Cross-Validation: K-Fold Cross-Validation provides a robust estimate of model performance by averaging results across multiple data partitions. The table presents three evaluation metrics for each regression model:

- **K-Fold MSE** – Mean Squared Error (lower is better)
- **K-Fold MAE** – Mean Absolute Error (lower is better)
- **K-Fold R^2** – Coefficient of Determination (higher is better)

Table 3: K-Fold Cross-Validation Performance Metrics

Model	K-Fold MSE	K-Fold MAE	K-Fold R^2
Linear Regression	25872.112	118.659	0.724
Polynomial Regression (Degree 2)	27329.288	106.807	0.716
Polynomial Regression (Degree 3)	127306.135	146.799	-0.300
Polynomial Regression (Degree 4)	18388500.109	611.440	-205.582
Ridge Regression	25872.083	118.659	0.724
Lasso Regression	25872.109	118.659	0.724
Elastic Net Regression	25872.023	118.659	0.724

Interpretation

- **Polynomial models (Degree $\geq$ 3)** showed dramatic overfitting, as indicated by negative R^2 values, proving that adding complexity did not capture meaningful relationships in the data.
- **Linear Regression and its regularized variants** (Ridge, Lasso, Elastic Net) provided consistent and accurate predictions, confirming that the underlying data pattern is essentially linear.
- The lack of improvement from regularization implies that multicollinearity and noise are minimal, and the model already generalizes well.
- **Polynomial Regression (Degree 2)** gave marginal improvement in MAE but a slightly worse overall fit, suggesting minor non-linearity — though not significant enough to justify increased model complexity.

Best Overall Model Based on K-Fold Validation

Linear Regression (and equivalent Ridge, Lasso, Elastic Net)

K-Fold Metrics:

$$\text{MSE} \approx 25{,}872, \quad \text{MAE} \approx 118.66, \quad R^2 \approx 0.724$$

These models demonstrate the most balanced and reliable performance across all folds. Polynomial models, particularly of Degree 3 and 4, exhibit clear overfitting, while the regularization methods confirm the model's stability rather than yielding any significant performance improvement.

4.Analysis Based on Test Error

Test Results: The test results evaluate each model's performance on unseen data after training and validation. The metrics considered are:

- **Test MSE (Mean Squared Error):** Measures the average squared difference between predicted and actual values.
- **Test MAE (Mean Absolute Error):** Represents the average magnitude of prediction errors.

- **Test R^2 (Coefficient of Determination):** Indicates how well the model explains the variance in target data.

A lower MSE/MAE and higher R^2 indicate better generalization and predictive performance.

Table 4: Test Error Metrics for Different Regression Models

Model	Test MSE	Test MAE	Test R^2
Linear Regression	17119.271	106.296	0.731
Polynomial Regression (Degree 2)	15730.499	92.004	0.753
Polynomial Regression (Degree 3)	43949.103	120.343	0.309
Polynomial Regression (Degree 4)	420047.038	255.838	-5.604
Batch Gradient Descent	14903.545	100.939	0.766
Stochastic Gradient Descent	17079.404	106.201	0.731
Mini-Batch Gradient Descent	16374.534	105.131	0.743
Polynomial Ridge Regression (Degree 2)	17119.249	106.296	0.731
Polynomial Lasso Regression (Degree 2)	17119.269	106.296	0.731
Elastic Net Regression	17119.258	106.296	0.731
Normal Equation	15701.960	102.553	0.753
SVD	15701.960	102.553	0.753

Interpretation

- **Linear Regression and Gradient Descent variants** performed well, confirming that the dataset follows a predominantly linear relationship.
- **Polynomial Regression (Degree 2)** offered a small but meaningful improvement, suggesting slight non-linear dependencies in the data.
- **Higher-degree polynomials (3 and 4)** drastically overfit the training data, leading to poor generalization and significantly higher test errors.
- **Regularized models** (Ridge, Lasso, Elastic Net) did not improve accuracy, indicating that the features were already well-scaled and not highly correlated.
- **Normal Equation and SVD** emerged as the most effective approaches, providing exact least-squares solutions with the highest precision and lowest error across all evaluation metrics.

Best Overall Models Based on Test error

Normal Equation and SVD Regression are the best model

Test Metrics:

$$\text{MSE} \approx 15{,}701, \quad \text{MAE} \approx 102.55, \quad R^2 \approx 0.753$$

These models generalize the best, achieving both high accuracy and strong stability across datasets. **Polynomial Regression (Degree 2)** provides a comparable alternative when mild non-linear behavior is present, whereas higher-degree and overly complex models exhibit clear overfitting.

Regularized and gradient-based models, though consistent and reliable, do not surpass the performance of the direct analytical methods such as the Normal Equation and SVD.

5.Analysis Based on Computational Time

Training Time reflects the computational efficiency of each regression model. It depends on the complexity of the algorithm, number of parameters, and amount of matrix computation or iteration performed during training. Lower training time indicates faster convergence and computational efficiency, while higher values suggest heavier computation or slower optimization.

Table 5: Training Time for Different Regression Models

Model	**Training Time (seconds)**
Linear Regression	0.002824
Polynomial Regression (Degree 2)	0.008017
Polynomial Regression (Degree 3)	0.007479
Polynomial Regression (Degree 4)	0.015069
Batch Gradient Descent	0.084035
Stochastic Gradient Descent	0.010179
Mini-Batch Gradient Descent	1.580085
Ridge Regression	0.003038
Lasso Regression	0.002583
Elastic Net Regression	0.001530
Normal Equation	0.001411
SVD	0.000479

Interpretation: The SVD and Normal Equation methods outperform others in computational efficiency and are ideal for smaller datasets where matrix inversion is feasible. Linear, Ridge, and Lasso regressions also train very quickly, benefiting from optimized linear solvers. Polynomial Regression's time grows exponentially with polynomial degree, making higher-order models computationally expensive and less scalable. Gradient Descent-based methods (Batch, Stochastic, Mini-Batch) require multiple iterations and are slower — but they're essential for large-scale datasets where matrix inversion becomes infeasible. Mini-Batch Gradient Descent, though the slowest, offers a good trade-off between stability and convergence in larger data applications.

Conclusion

- **Fastest Model:** Singular Value Decomposition (SVD) — **0.000479 s**
- **Next Fastest Models:** Normal Equation and Elastic Net Regression
- **Most Computationally Intensive:** Mini-Batch Gradient Descent (**1.580 s**)

Analytical models such as **SVD** and the **Normal Equation** are ideal for small to medium datasets due to their superior speed and numerical precision. Iterative models (especially **Batch** and **Mini-Batch Gradient Descent**) are comparatively slower but offer better scalability for large datasets, where matrix inversion becomes computationally expensive.

Furthermore, **Polynomial Regression** exhibits non-linear growth in training time with increasing degree, confirming that higher model complexity leads to greater computational cost without proportional accuracy gains.

5 Conclusion

The following key conclusions were drawn from the **Mobile Price Prediction** project:

- The project successfully developed and evaluated multiple regression models, including **Linear**, **Polynomial**, **Regularized**, and **Gradient Descent**–based methods, to predict mobile phone prices from specifications such as *RAM, battery capacity, screen size, and camera quality*.
- The dataset exhibited **predominantly linear relationships**, as linear and regularized models consistently outperformed higher-degree polynomial models.
- The **Normal Equation** and **Singular Value Decomposition (SVD)** methods achieved the **best overall performance**, with:
  - Mean Squared Error (MSE): $\approx 15{,}701$
  - Mean Absolute Error (MAE): ≈ 102.55
  - Coefficient of Determination (R^2): ≈ 0.753
- These analytical methods provided both **high accuracy** and **computational efficiency**, making them ideal for *small to medium-sized datasets*.
- **Linear Regression** and its regularized forms (**Ridge, Lasso, Elastic Net**) produced nearly identical results, confirming **model robustness** and indicating **minimal multicollinearity** in the dataset.
- **Polynomial Regression** of higher degrees (3 and 4) showed **significant overfitting**, while the second-degree model captured minor non-linearities but without notable performance improvement.
- **Gradient Descent**–based models converged to similar results as Linear Regression, verifying their correct implementation but demonstrating slower performance due to iterative optimization.
- Overall, **simpler linear models**, particularly the **Normal Equation** and **SVD Regression**, offered the best trade-off between **accuracy, generalization, and computational efficiency**.
- The study emphasizes the importance of **model interpretability** and **computational feasibility** over unnecessary model complexity, providing valuable insights for future predictive modeling tasks in *consumer electronics pricing* and related domains.

References

1. Kaggle. *Mobile Price Prediction Dataset*. Available at: `https://www.kaggle.com` [Accessed November 2025].
2. Kevin Patrick Murphy. *Probabilistic Machine Learning: An Introduction*. MIT Press, 2022.