**Mobile Price Trends: A Data-Drivenpasted-image.tiff Predictive Study**

*Vaidyula RajaRohan Reddy,*

*2203A52126,*

*2203a52126@sru.edu.in,*

*SR University.*

**Abstract :**

Most of the people in the society don’t know the exact price of the mobiles they simply believe in the company and the seller.If a person want to buy a mobile he don’t know the actual price of the mobile on the basis of it’s features.This project is for to apply machine learning methods to predict the mobile prices with the given features.

In this we are going to do implement machine learning Algorithms such as linear regression , Support Vector Machine , Lasso regression , Ridge regression and KNN regressor. And then finding mean squared error , mean absolute error , root mean squared error than by bootstrapping method we will select the best model with low uncertainty.

The main intention of this project is to provide the customer the valuable price prediction and make purchase or not based on predicted price of ML model and price claimed by the company or the seller.It will help the costumer to think and buy the mobile and see the mobile price is worth it or not.

**Key words :**

Linear regression , Support Vector Machine , Lasso regression , Ridge regression , KNN regressor , mean squared error , mean absolute error , root mean squared error , Bootstrapping , Best model.

|  |  |  |
| --- | --- | --- |
|  |  | Pg no. |
| **Introduction** |  | 3 |
| **Literature Review** |  | 3 |
| **Data Methodology** |  | 4 |
| **Results** |  | 8 |
| **conclusion** |  | 16 |
| **References** |  | 18 |

**Introduction :**

From past many years mobile market is evolving continuously with new models and predicting or forecasting the price of the mobile is been a task for customer and mobile makers. Customer and producer of mobile are ready to know what influence the prices of mobiles. This project will help the both user and the manufactures.

The main activity of this project is to Collect data , data preprocessing . This will depend on features of the mobile and their numerical values corresponding to the prices. To build an accurate error free predictive model we use general Machine learning algorithms . Here , the price of the mobile is the target variable and it is continuous data so we can apply regression model on it . Our main intention is to make every one know in depth understanding of mobile prices .

we make sure that which model is the best model for the data by calculating the Mean square error (MSE) , Mean absolute error (MAE) , Root mean square error (RMSE) , Bootstrapping and Principal component analysis (PCA).

**Literature Review :**

The following literature review on prediction of prices of Mobiles using machine learning would include the different methods from research. Here's an example of the literature review of the our topic called prediction of the Mobile prices using the machine learning models.

The prediction from ML models has been widely increased in the modern world. This literature review plays a vital role in the studies involving in this field.

*2.1* ***ML in price prediction :***

*Machine learning is very important in price prediction of mobiles. By machine learning we can do many things such as predicting , classifying , clustering etc. Researchers had made many models for predicting prices like linear regression , decision tree , random forest , deep learning . And these algorithms done very good job till date.*

*2.2* ***Data Sources :***

*Here , we collected data form kaggale but researchers have took their data from various platforms and some of them collected the data manually by checking the price of each mobile . They took data from marketing websites and product specifications . In ML models data Quantity ,Quality and data preprocessing plays a crucial role in the prediction and efficiency of the model .*

*2.3* ***Model Metrics :***

*The power of the ML model is denoted by many metrics like Bootstrapping ,MSE , MAE , RMSE .This methods will provide the how well the model was performing in predicting mobile prices.*

*2.4* ***Future Engineering :***

*Researchers have done a great job in finding the innovative approaches to find meaning full features to predict the price. The features are Weight , Resolution , CPU frequency , Internal Memory , RAM , Camera , Battery etc.*

*2.5* ***Challenges and Research Gaps :***

*Several barriers to mobile pricing prediction , including market , frequently product launching , and client preferences. The best-performing ML algorithms for diverse settings are being identified and addressed in continuing research.*

*2.6* ***Applications***

*Not only research , ML based price prediction has found in many fields such as E - commerce , retail , and investment etc.*

**Data Methodology :**

*3.1****. Data Description:***

*The price of a smartphone is determined by several aspects, including resolution, brand, size, weight, image quality, RAM, battery, and CPU power. Using the characteristics listed above, we hope to estimate the pricing of mobile phones in this dataset.*

*1>Product\_id: This is a unique identification assigned to each mobile phone in the collection. It does not immediately contribute to price prediction, but it can assist in tracking and identifying certain gadgets.*

*2>Sale: If the phone is now on sale or has a discount, this feature may indicate that. Because sales and discounts may influence the price of a phone, this function may be useful for price prediction.*

*3>Weight: The mobile phone's weight in grams. Heavier phones may have larger batteries or more features, which might affect their price.*

*4>Resolution: This most likely refers to the phone's display screen resolution, which is normally indicated in pixels (e.g., 1920x1080). Phones with higher screen resolutions are frequently more expensive since they provide greater display quality.*

*5>PPI (Pixels Per Inch): PPI is a metric for determining pixel density on a screen. Phones with higher PPI values have clearer and more detailed screens, which may influence their pricing.*

*6>CPU Frequency: The central processing unit (CPU) clock frequency, commonly measured in gigahertz (GHz). Faster CPUs can perform more demanding jobs, but they may be more expensive.*

*7>Internal Memory: This refers to the phone's built-in storage capacity, which is commonly measured in gigabytes (GB) or terabytes (TB). Phones with bigger internal storage capabilities are sometimes more expensive.*

*8>RAM (Random Access Memory): RAM is the phone's temporary memory for executing programs. Phones with more RAM tend to operate better and may be more expensive.*

*9>Back Camera:The quality of the rear camera has a significant impact on the price of a smartphone. This feature may specify the resolution of the camera in megapixels or other camera-related information.*

*10>Front Camera: As with the rear camera, the quality and specifications of the front-facing camera can influence the price, particularly for devices geared toward selfie photography.*

*11>Battery: This feature most likely specifies the phone's battery capacity, which is typically measured in milliampere-hours (mAh). Phones with larger batteries can provide longer battery life and may be more expensive.*

*12>Thickness: The thickness of the phone, which is commonly measured in millimeters (mm). Slimmer phones may have a more appealing design, but they may sacrifice battery capacity or other features.*

*3.2.* ***Data Analysis :***

*These graphs help you understand how individual features contribute to the regression algorithm. Features with different, non-overlapping distributions are typically more interactive and informative for regression. Featuresl may not be as useful for differentiating between analysing these histograms, you can make informed decisions about feature selection, model and engineering to improve the efficiency and the performance of your model.*

*3.3* ***Correlation Matrix :***

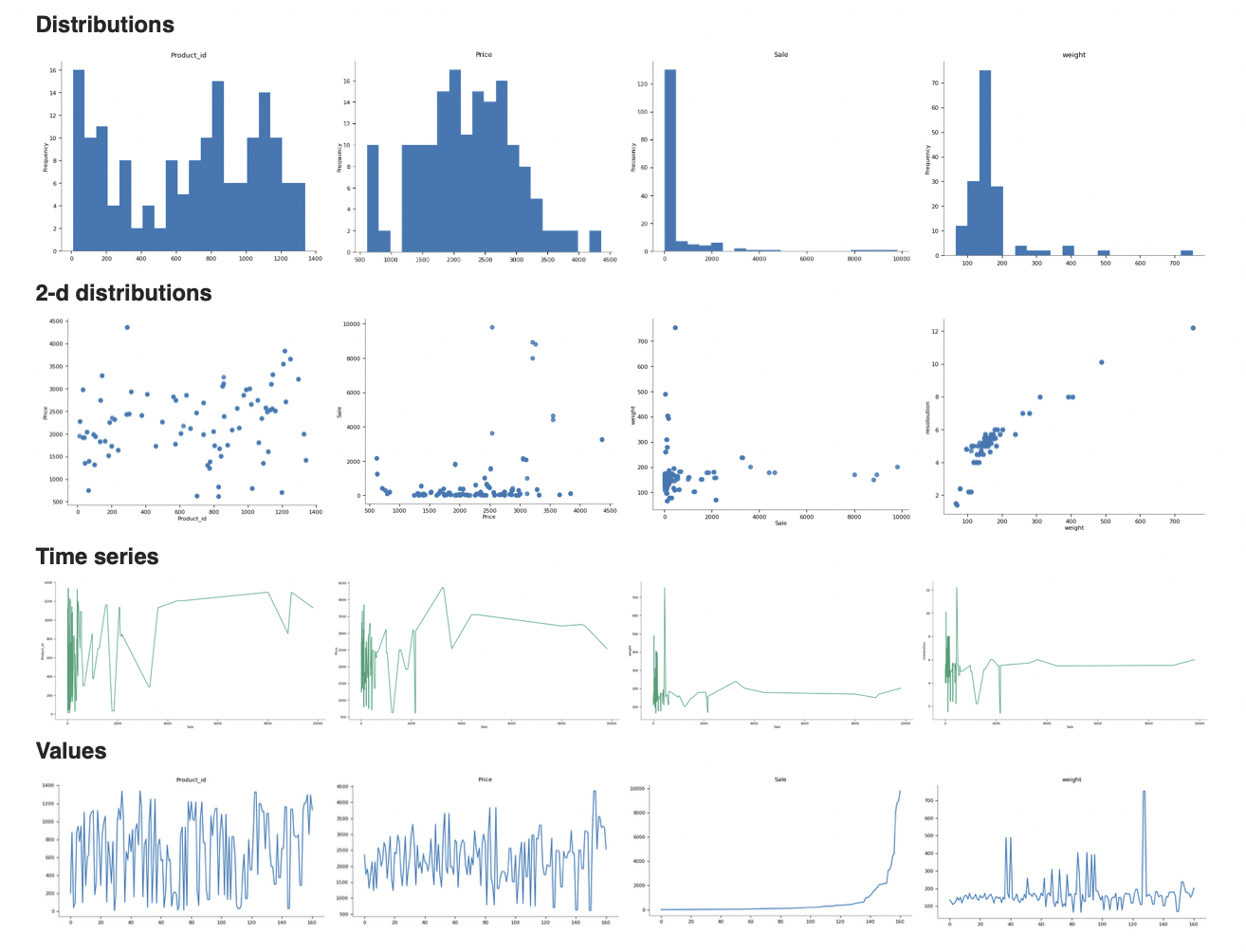
*Correlation matrix is a table that gives us information about the correlation coefficients between many features .The cell in the matrix shows the correlation between two features. The correlation coefficient, denoted as "r," ranges from -1 to 1,*

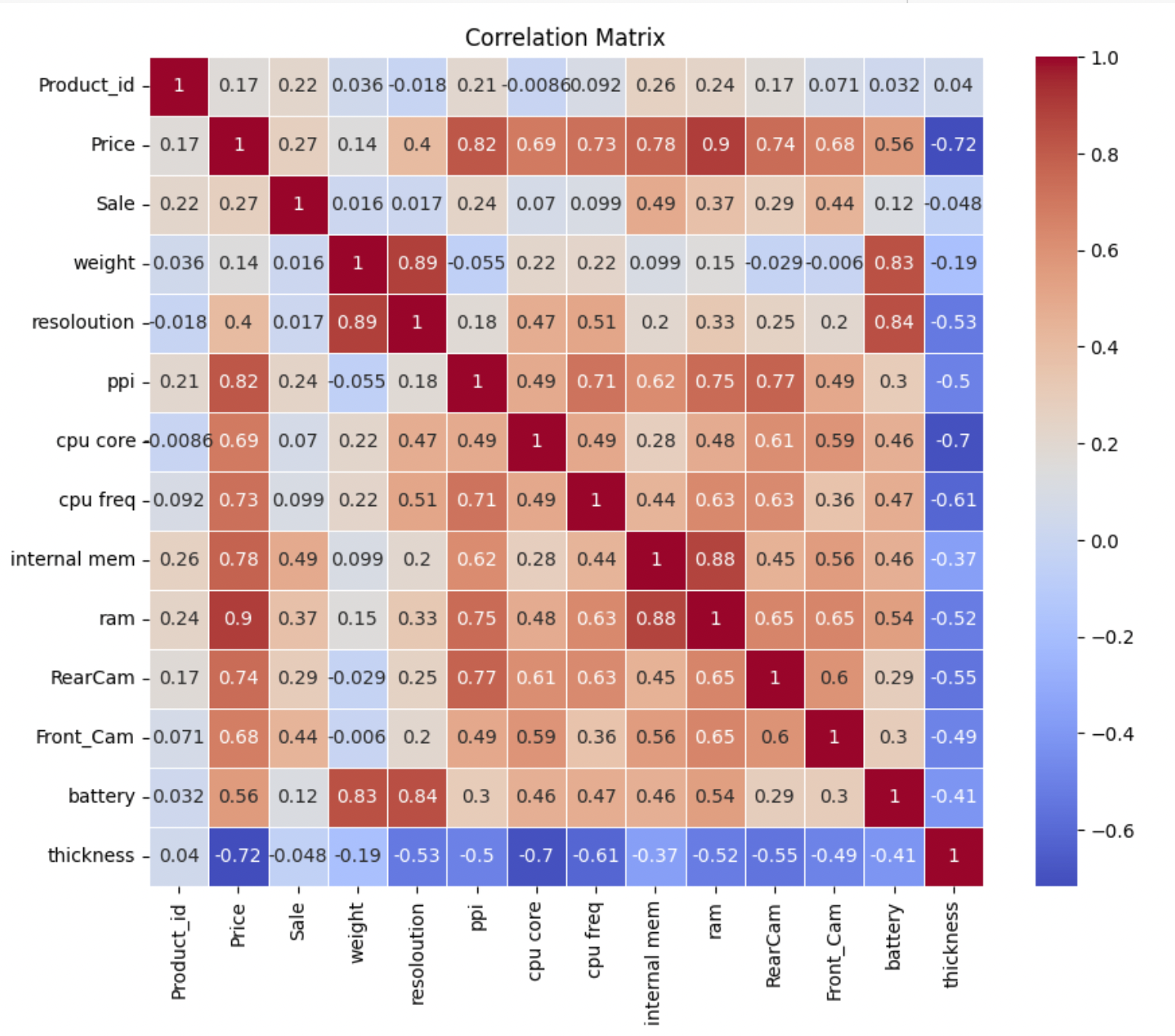
*where:*

*1 indicates a perfect positive correlation.*

*0 indicates no correlation.*

*--1 indicates a perfect negative correlation.*





*3.4.* ***Data Splitting:***

*Train-Validation-Test Split: Divide the dataset into training, validation, and test data sets. The training set is used to train the model, the validation set is used to train the machine learning models and the test set is used to evaluate model performance and accuracy.*

*Model Selection We selected five machine learning models for our analysis: Linear Regression, Support Vector Machines, Lasso, Ridge regression and K-Nearest Neighbours. These models were chosen based on their suitability for binary classification tasks*

*3.5.* ***Experimental Setup :***

X\_train, X\_test, y\_train, y\_test = train\_test\_split(features, target, test\_size=0.2, random\_state=42)

*This code performs the actual data split. It takes the shuffled Data Frame, which contains all the data, and splits it into two separate Data Frames: the training set and the test set. The proportions of the splits are determined by the fractions (80 percent for training, and 20 percent for testing) as specified in the indexing, but these proportions can be adjusted based on the specific requirements of your machine learning task.*

**Results**

***4.1. Bootstrapping***

*Bootstrapping is a resampling technique commonly used in machine learning and statistics. It involves repeatedly sampling data from your dataset with replacement to create various new datasets, each of the same size as the size of the original. The intention of bootstrapping is to create various new datasets, each of the same size as the original. These datasets are called "bootstrap samples. Bootstrapping helps in assessing the variability and also the efficiency of your model. By training multiple machine learning models on different bootstrap samples, you can evaluate how well your trained model is going to adapt to various subdomains of the data.*

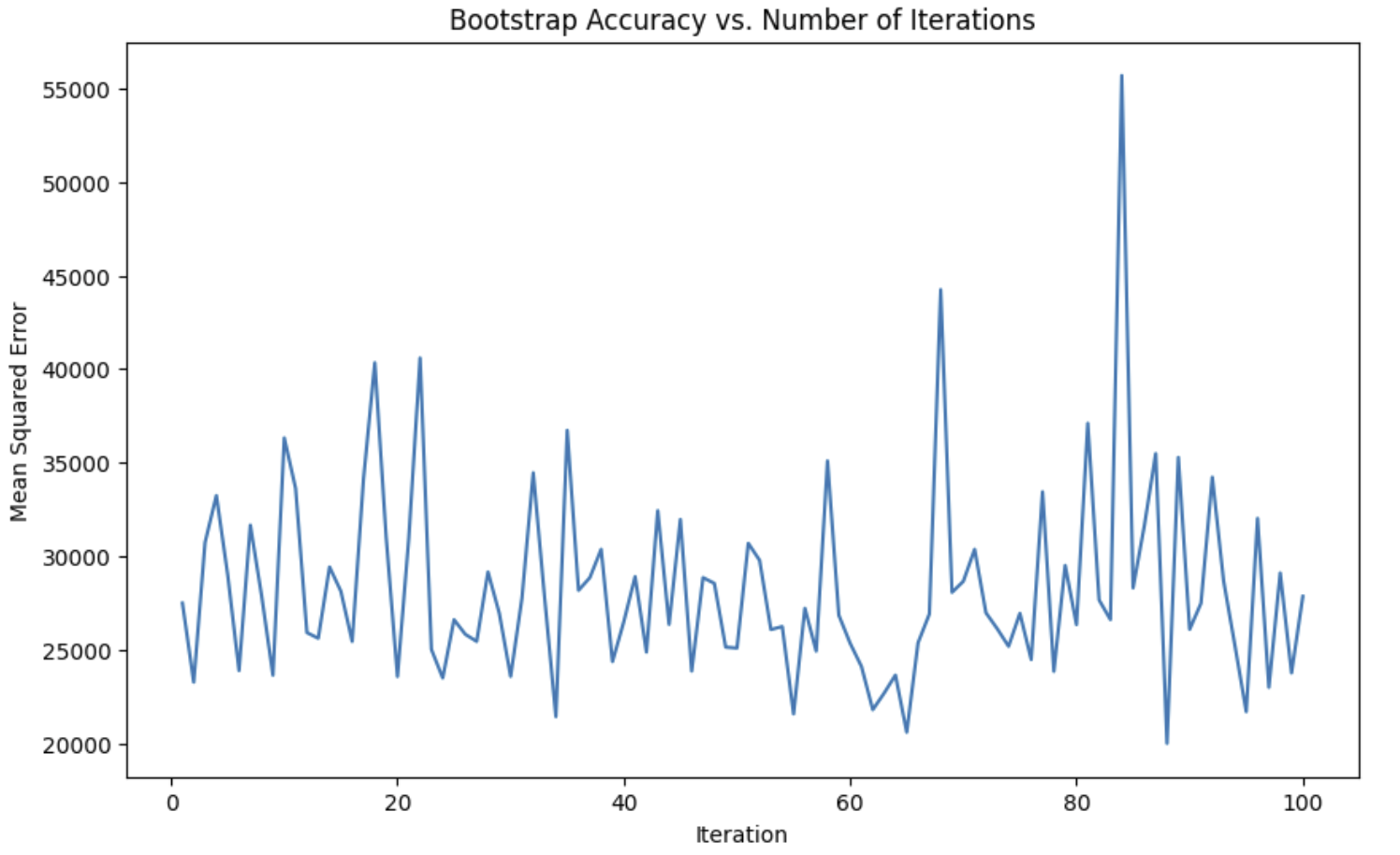
*Bootstrapping can be used to estimate confidence intervals for various performance metrics of your model, such as accuracy or mean squared error. By repeatedly resampling the data and evaluating the model on each sample, you can get a distribution of the metric and compute its confidence interval.*

*4.1.1****. Linear regression model***

*Linear regression is a statistical method used for convey the relationship between a target variable and one or more feature variables by fitting the best fit line with a linear equation or the best fit line to the observed data in given dataset. Linear regression is very simple and mostly-used technique for predictive machine learning modelling, helping to make predictions or understand the strength of relationships between variables.*

*Mean squared error: 28947.31450*

*Standard Deviation: 23062.22989175004*

* + 1. ***Mean Squared Error (MSE):*** *It is a generally statical measure used to measure the average squared difference between the actual values and the predicted values in a regression or machine learning prediction model. It generally conveys us about the model's accuracy and its efficiency, with lower MSE i.e mean squared error indicating a better or the best fit line for our trained model. MSE is calculated by averaging the squares of the estimated errors, making more errors contribute more to the overall score. It's a important tool for calculating and comparing the performance of various models in statistical and machine learning implementation.*
    2. ***Standard deviation:*** *It is a statistical measure of the amount of variation in a set of the data given in the dataset . It gives how much individual data points will deviate from the average of the data set. A higher standard deviation(s.d) indicates greater changes, while a lower standard deviation(s.d) will indicate that the data points are closer to the mean provided. It is commonly used in various fields and technology, including statistics, finance, and science, to understand and describe the spread of the dataset of the forest fires within our dataset.*

*bootstrap accuracies vs iterations graph*

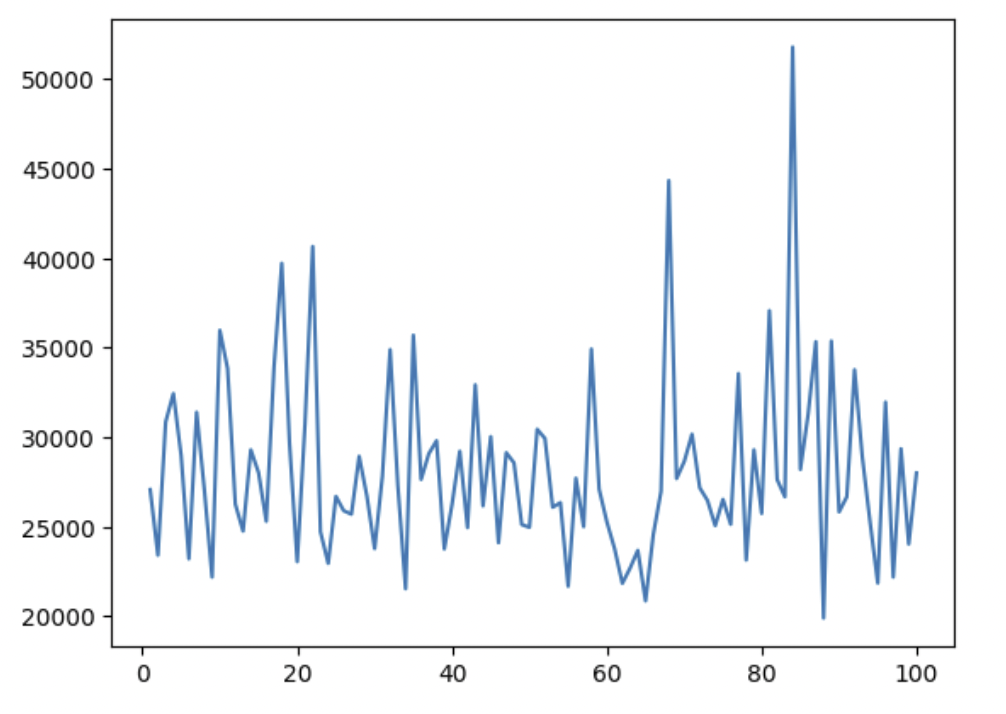
*4.1.2****. Ridge regression model***

*Ridge regression is a regularization technique used in linear regression to prevent overfitting and improve model stability. It adds a penalty term to the standard linear regression equation, which is the L2 regularization term. This penalty term discourages large coefficients for the predictor variables, effectively shrinking them towards zero. By doing so, ridge regression reduces the model's sensitivity to individual data points and multicollinearity, making it more robust and generalizable. It's especially useful when dealing with high-dimensional datasets or when multicollinearity is a concern. The strength of regularization is controlled by a hyperparameter called lambda, which can be tuned to optimize model performance.*

*Mean squared error: 29924.646210855382*

*Standard Deviation: 23107.47409092388*

* + 1. ***Mean Squared Error (MSE):*** *It is a generally statical measure used to measure the average squared difference between the actual values and the predicted values in a regression or machine learning prediction model. It generally conveys us about the model's accuracy and its efficiency, with lower MSE i.e mean squared error indicating a better or the best fit line for our trained model. MSE is calculated by averaging the squares of the estimated errors, making more errors contribute more to the overall score. It's a important tool for calculating and comparing the performance of various models in statistical and machine learning implementation.*
    2. ***Standard deviation****It is a statistical measure of the amount of variation in a set of the data given in the dataset . It gives how much individual data points will deviate from the average of the data set. A higher standard deviation (s.d) indicates greater changes, while a lower standard deviation (s.d) will indicate that the data points are closer to the mean provided. It is commonly used in various fields and technology, including statistics, finance, and science, to understand and describe the spread of the dataset of the forest fires within our dataset.*

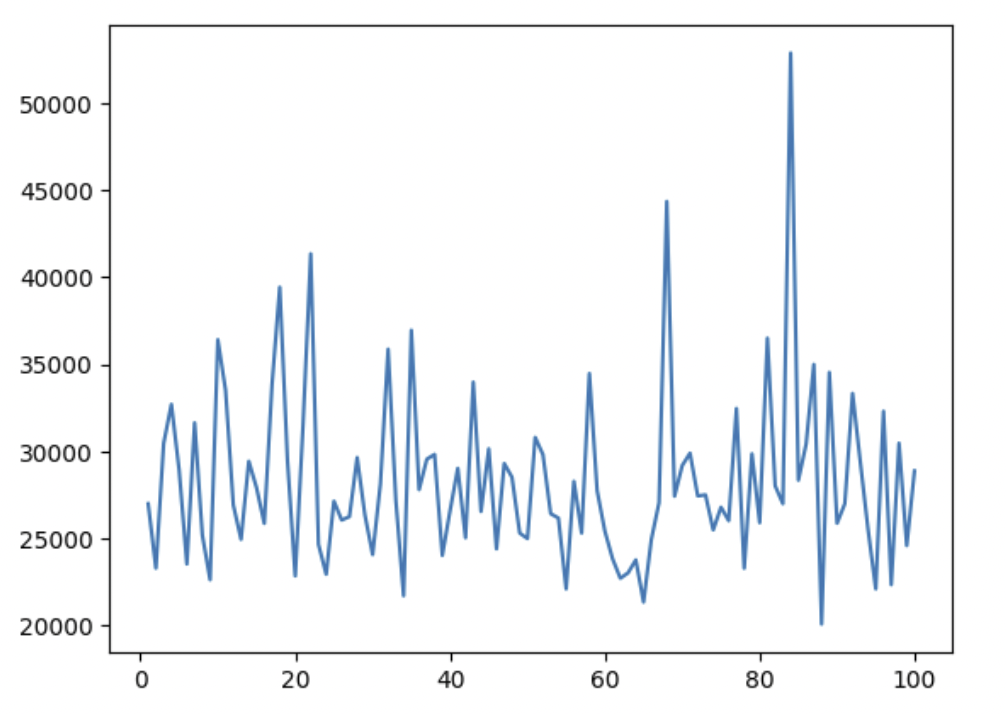
*bootstrap accuracies vs iterations graph*

*4.1.3****. Lasso regression model***

*Lasso regression, short for "Least Absolute Shrinkage and Selection Operator," is a type of linear regression that adds a penalty term to the ordinary least squares (OLS) objective function. This penalty term encourages sparsity in the regression coefficients by adding the absolute values of the coefficients to the loss function. It is commonly used for feature selection, as it tends to drive some coefficients to exactly zero, effectively eliminating those features from the model. Lasso regression can help prevent overfitting and improve model interpretability by reducing the complexity of the model. The degree of regularization in lasso is controlled by the regularization hyperparameter, often denoted as "alpha."*

*Mean squared error: 23078.857965969488*

*Standard Deviation: 23352.977933069393*

* + 1. ***Mean Squared Error (MSE):*** *It is a generally statical measure used to measure the average squared difference between the actual values and the predicted values in a regression or machine learning prediction model. It generally conveys us about the model's accuracy and its efficiency, with lower MSE i.e mean squared error indicating a better or the best fit line for our trained model. MSE is calculated by averaging the squares of the estimated errors, making more errors contribute more to the overall score. It's a important tool for calculating and comparing the performance of various models in statistical and machine learning implementation.*
    2. ***Standard deviation:*** *It is a statistical measure of the amount of variation in a set of the data given in the dataset . It gives how much individual data points will deviate from the average of the data set. A higher standard deviation(s.d) indicates greater changes, while a lower standard deviation(s.d) will indicate that the data points are closer to the mean provided. It is commonly used in various fields and technology, including statistics, finance, and science, to understand and describe the spread of the dataset of the forest fires within our dataset.*

*bootstrap accuracies vs iterations graph*

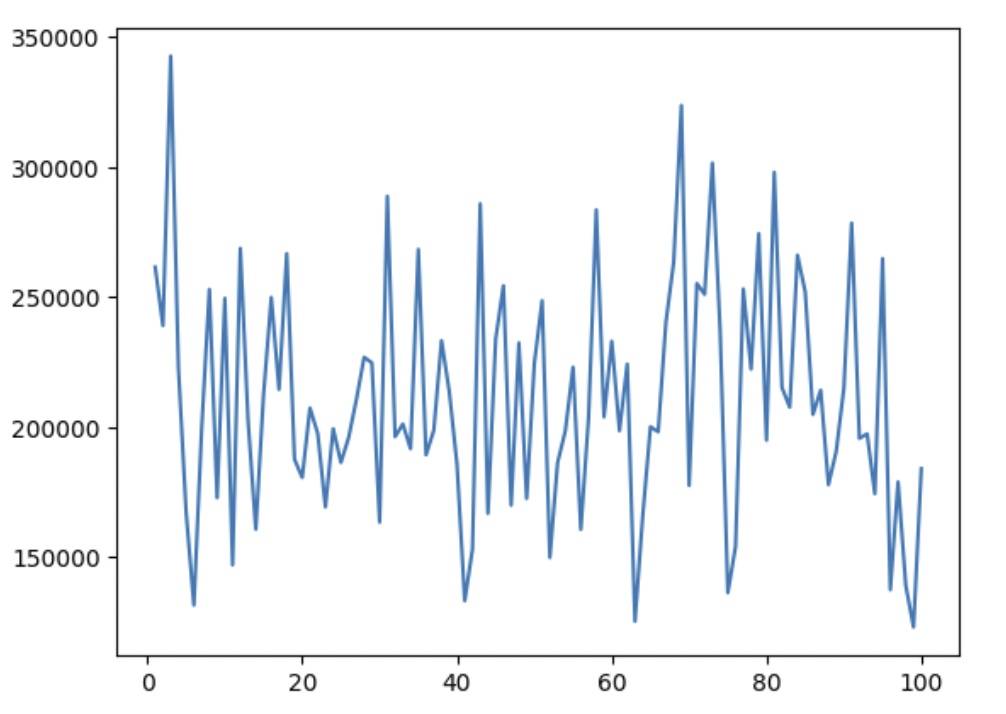
*4.1.4****. KNN regression***

*K-Nearest Neighbors (KNN) is a supervised machine learning algorithm used for classification and regression tasks. It is a simple and intuitive algorithm that can be used for various types of data analysis and pattern recognition tasks. Here's some information about KNN:*

*Mean squared error: 72447.85454545455*

*Standard Deviation: 157930.41696969696*

* + 1. ***Mean Squared Error (MSE):*** *It is a generally statical measure used to measure the average squared difference between the actual values and the predicted values in a regression or machine learning prediction model. It generally conveys us about the model's accuracy and its efficiency, with lower MSE i.e mean squared error indicating a better or the best fit line for our trained model. MSE is calculated by averaging the squares of the estimated errors, making more errors contribute more to the overall score. It's a important tool for calculating and comparing the performance of various models in statistical and machine learning implementation.*
    2. ***Standard deviation:*** *It is a statistical measure of the amount of variation in a set of the data given in the dataset . It gives how much individual data points will deviate from the average of the data set. A higher standard deviation(s.d) indicates greater changes, while a lower standard deviation(s.d) will indicate that the data points are closer to the mean provided. It is commonly used in various fields and technology, including statistics, finance, and science, to understand and describe the spread of the dataset of the forest fires within our dataset*



*bootstrap accuracies vs iterations graph*

*4.1.5.* ***Support vector machine (SVM):***

*Support Vector Machine (SVM) is a powerful and widely used machine learning algorithm for classification, regression, and outlier detection. It's particularly popular in both academic and practical applications due to its effectiveness and versatility. Here's some key information about SVM is a versatile and powerful algorithm, but it requires careful parameter tuning and consideration of the specific problem you are trying to solve. When used appropriately, SVM can deliver excellent results in a wide range of machine learning tasks.*

*Mean squared error: 598064.1619885144*

*Standard Deviation: 573245.4156957181*

* + 1. ***Mean Squared Error (MSE):*** *It is a generally statical measure used to measure the average squared difference between the actual values and the predicted values in a regression or machine learning prediction model. It generally conveys us about the model's accuracy and its efficiency, with lower MSE i.e mean squared error indicating a better or the best fit line for our trained model. MSE is calculated by averaging the squares of the estimated errors, making more errors contribute more to the overall score. It's a important tool for calculating and comparing the performance of various models in statistical and machine learning implementation.*
    2. ***Standard deviation:*** *It is a statistical measure of the amount of variation in a set of the data given in the dataset . It gives how much individual data points will deviate from the average of the data set. A higher standard deviation(s.d) indicates greater changes, while a lower standard deviation(s.d) will indicate that the data points are closer to the mean provided. It is commonly used in various fields and technology, including statistics, finance, and science, to understand and describe the spread of the dataset of the forest fires within our dataset*

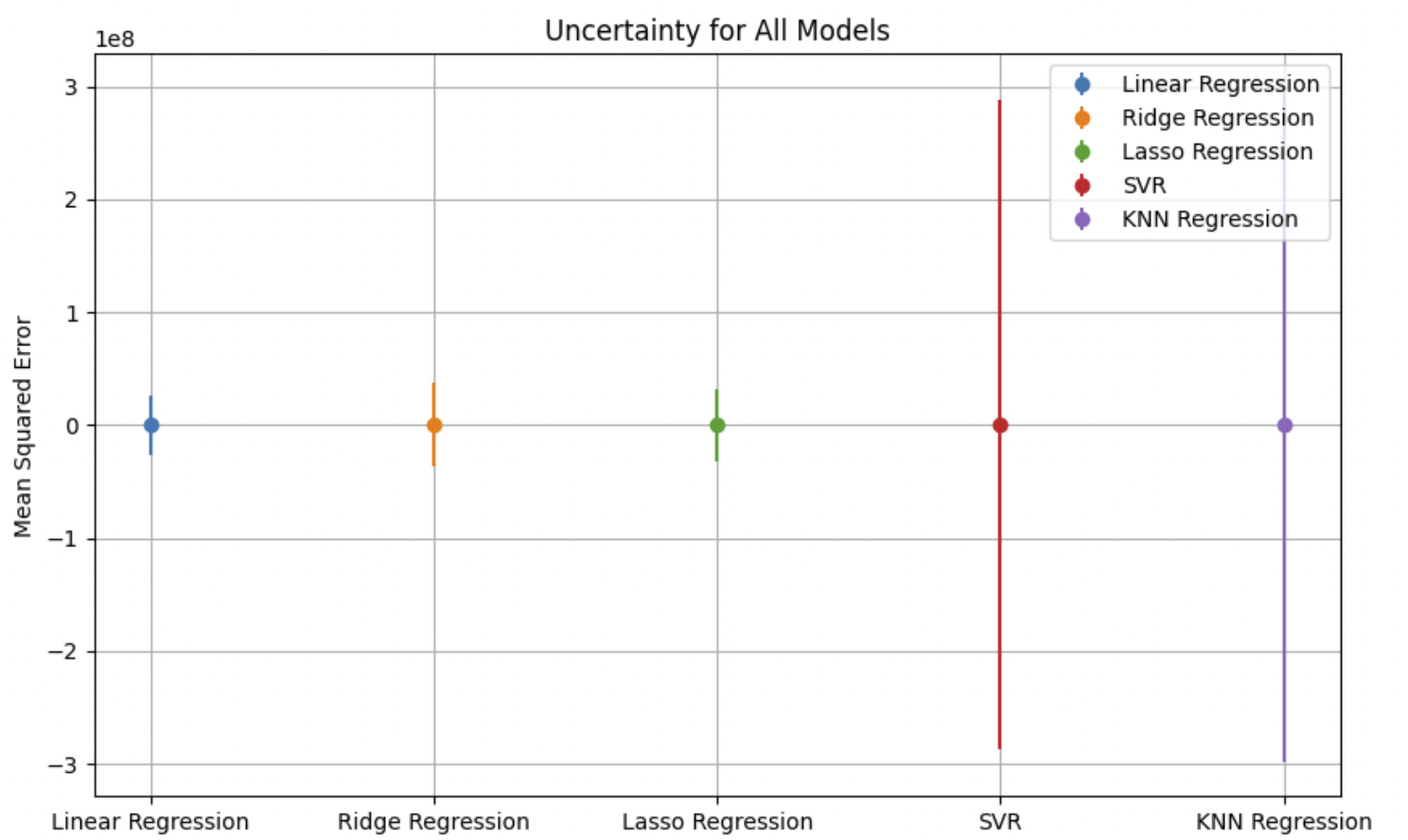
*bootstrap accuracies vs iterations graph*

**conclusion:**

***Interpretation****: The Linear regression model performed the best overall, achieving the lowest mean squared error (MSE) with the value of 28947.31450052959 and also the lower standard deviation with value of 23062.22989175004. Both the linear and KNN Regression models had somewhat low mse and low standard deviation which also make them best fit models. But comparatively lasso is best fit model among all the other machine learning models applied.*

|  |  |  |
| --- | --- | --- |
| Model | Mean squared error | Standard deviation |
| linear | *28947.3145005296* | *23062.22989175* |
| knn | *72447.85454545455* | *157930.416969697* |
| lasso | *23078.857965969488* | *23352.977933069393* |
| ridge | *29924.646210855382* | *23107.4740909239* |
| SVM | *598064.16198* | *573245.415695718* |

*Uncertainty fir all models by bootstrapping*

*5.1.* ***Improvement :***

*One opportunity for future work lies in further sophistication of the implemented machine learning algorithm. Artificial neural networks could also provide a good frame-work for modelling this data. As we mentioned in our Exploratory Visualization section, no individual feature clearly distinguishes the gamma ray and hadron classes, and well- performing machine learning models must be able to capture interactions between features. There is really no rival to the neural network algorithm when it comes to automatically detecting and leveraging feature interactions. However, this benefit comes with the added cost of high risk of overfitting, long training times, and a wide array of hyperparameters to optimize. Given more time to complete this capstone, my focus would have moved to neural networks next.*

**References**

*1. \*\*"Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" by Aurélien Géron:\*\**

*- As mentioned earlier, this book is a practical guide to machine learning using popular Python libraries. It covers essential concepts and provides hands-on examples.*

*2. \*\*"Python Machine Learning" by Sebastian Raschka and Vahid Mirjalili:\*\**

*- This book explores machine learning using Python and includes topics like supervised and unsupervised learning, dimensionality reduction, and more.*

*3. \*\*"Machine Learning Yearning" by Andrew Ng:\*\**

*- Authored by the co-founder of Coursera and a prominent figure in the field of machine learning, this book focuses on the practical aspects of deploying machine learning systems and managing ML projects effectively.*

*4. \*\*"Introduction to Machine Learning with Python: A Guide for Data Scientists" by Andreas C. Müller and Sarah Guido:\*\**

*- This book is a great introduction to machine learning with a focus on practical examples using the Scikit-Learn library. It's suitable for beginners and intermediate learners.*

*5. \*\*"Deep Learning" by Ian Goodfellow, Yoshua Bengio, and Aaron Courville:\*\**

*- If you're interested in delving into deep learning, this comprehensive book is a key resource. It covers the theoretical foundations of deep learning and provides insights into advanced neural network architectures.*