A Statistical Method under Missing Data Framework for Car Price Prediction

Yanjiao Yang (1177963), Hanyue Liang (1179743), Huiying Lin (1199232), Ruixin Pang (1209150), Wenxin Song (1183796)

April 26, 2023

Novel Approach under Missing Data Framework for Car Price Prediction Proposed by Duke University Students

April 26, 2023 - Used car price prediction is a popular research topic in the field of machine learning and data analytics. The goal of used car price prediction is to build a model that can accurately predict the selling price of a used car based on a set of input features such as the car's make, model, year, mileage, and other relevant factors. Accurate used car price prediction can be helpful for both buyers and sellers in determining a fair market value for a used car.

Like all other real-world prediction problems, one of the key challenges in used-car price prediction is dealing with missing data. It is common for some of the input features to be missing or incomplete, which can make it difficult to build an accurate prediction model. To address this challenge, a team of students from Duke University proposed a novel approach that combines regression model with missing data to improve the accuracy and reliability of the price prediction model. Traditional techniques for handling missing data generally involve data imputation such as mean or median imputation, regression imputation, and multiple imputation. The students developed a Lasso-missing technique, and leveraged ensemble methods including voting regressor and meta-analysis to improve price prediction.

The highlight of their proposed method is an improved Lasso technique within a missing data framework to enable automatic variable selection and to identify the key variables that impact car prices, together with further ensembling methods to leverage the strengths of multiple models. Different from the standard Lasso regression which needs to be performed on the complete data set, the team of students proposed a method that incorporates missingness into the model design, which treats missing values as a component of optimization and avoids introducing bias from simply discarding or imputing the missing data. This strategy improves the accuracy and reliability of the predictions for second-hand car prices in North Carolina. To further improve the prediction performance, the team incorporated a voting regressor and meta-analysis into their model framework. The voting regressor is an ensemble learning technique that combines the predictions of multiple individual regression models to obtain a final prediction. The models used in voting regressor were selected by applying several basic models to the complete data set obtained from data imputation. Overall, the proposed statistical model offers both theoretical and applicable advantages, including performing automatic variable selection, providing a more comprehensive understanding of the relationships between the available features and car prices, enhancing model performance through ensembling methods, and enabling generalization to other prediction tasks.

The students' novel approach has proven to be effective in predicting used car prices in North Carolina. In the future, the team plans to release a software with their algorithm encapsulated.

With the software, the algorithm will be more user-friendly and can be applied to perform predictions under broader scenarios. The students state that the proposed work aims at reducing information asymmetry and leading to greater customer satisfaction and business outcomes by providing more accurate and reliable pricing information to buyers and sellers. Additionally, the algorithm developed through the project has broader applications in other industries and domains where predictive modeling is used to inform decision-making.

Overall, the proposed method that integrates missing data into model design and ensemble models is a promising approach to predicting used car prices. The students believe that it can be extended to other prediction tasks of various fields. By bridging human experience and intuition with modern advances in computer science and statistics, the students hope that their work will contribute to the growing trend of integrating big data-based approaches into various fields.

Frequently Asked Questions

1. Can you briefly introduce your algorithm and proposed work?

We aim to provide a statistical method that integrates missing data into model design to help users better deal with missingness and make better predictions. Specifically, we first propose an improved Lasso regression. Lasso regression is a common model and has the advantage of variable selection which is helpful for high-dimensional data, but it does not consider missing values. Based on this limitation, we improve the standard Lasso regression by incorporating missingness into model design. Basically, the method takes the missing rate into consideration and solves a nonconvex minimization problem by using the soft thresholding and Euclidean projection onto a ℓ_1 ball in each iteration. Further, we use voting regressor and meta analysis to combine our improved Lasso model with other models to obtain a more reliable prediction result. We apply our algorithm to used car data in North Carolina to predict the car prices, which brings satisfying improvement to the performance of the predicted model. This outcome supports the feasibility of the algorithm.

2. What is the improvement of your model design compared with existing work?

Our proposed statistical model provides several improvements compared to existing work. Firstly, we implemented an improved Lasso technique within a missing data framework, which enables automatic variable selection and helps identify the key variables that impact car prices based on an interpretable prediction model. Secondly, our approach provides a more comprehensive understanding of the relationships between the available features and car prices, taking into account the missing data in the dataset and improving the accuracy and reliability of our predictions for second-hand car prices in North Carolina. Thirdly, we incorporated voting regressor and meta-analysis into our prediction model framework, which further enhanced the prediction. In the future, our model may be generalized to other prediction tasks and is particularly useful when dealing with large amounts of missing data. We also address the non-convexity issue and provide the algorithm for the proposed model. In summary, in the used-car price problem background, our approach improves on existing work by incorporating missingness into the model design, providing a more accurate and comprehensive understanding of relationships, and incorporating advanced techniques for prediction.

3. What is the most important value your proposed work provides?

The proposed work provides both theoretical values and applicable values. On one hand, it combines standard Lasso with the missing data framework. As for real-world data set, dealing with missing data is an important step and a common challenge in data analysis. Common methods to deal with missing value include deletion and mean/median/mode imputation. Those methods are simple and easy to carry out, but sometimes arbitrary and cause bias. Rather than directly dropping or imputing the missing data, our method formulates a nonconvex optimization problem which takes missingness into consideration and perform variable selection. On the other hand, our algorithm can be encapsulated as a software for users to implement the method directly and obtain satisfying predictions. Users don't need to spend time and energy to write their own code, and the only thing that they need to do is to turn to our software and make some clicks. In brief, our software implements an improved Lasso, based on which the voting regressor and meta analysis are further applied to improve model performance and capture both linear and nonlinear relationship such that users will obtain satisfying predictions.

4. What preparation work do you do before proposing your model?

To prepare for proposing our model, we first conducted exploratory data analysis and cleaned the given dataset. This involved filtering the data to only include entries from North Carolina and dropping columns unlikely to significantly impact predicting car prices. We also handled outliers in the price column by removing prices greater than 1 million dollars and identifying remaining outliers using median absolute deviation. Rows with missing values for features with missing rates less than 4% were dropped, while the remaining features (condition, cylinders, drive, size, type, and paint color) had missing rates ranging from 20% to 65%. As the missing rate is high and the missing data in these features could contain valuable information, we investigated various data imputation or handling methods such as Lasso regression with automatic feature selection, KNN, mean, and median imputation. Based on these preparation works, we proposed a model combining the ensembled data imputation and prediction algorithms method with Voting and Meta-Analysis, which can handle missing values and make precise predictions in various missing data settings, including predicting used-car prices. More importantly, in our model, we innovatively proposed the Lasso-Missing method and combined it into the Voting and Meta-Analysis.

5. Can you provide more details on how your improved Lasso algorithm deals with missing data?

First, given the design matrix with missing data, we first formulate a minimizing problem based on standard Lasso. Our method considers the missing rate and can be regarded as a generalization of standard Lasso regression. Technical details can be obtained in the next section. One potential issue is that when there is a large amount of missing data, the minimization problem is not necessarily a convex optimization, which can be troublesome for implementation. To solve the minimization problem, we formulate and implement a composite version of projected gradient descent algorithm. Specifically, the update in each iteration makes use of soft thresholding and Euclidean projection onto the ℓ_1 ball. Soft thresholding is a method used to shrink the values of a vector towards zero by applying a penalty term proportional to the absolute value of the elements of the vector. Euclidean projection onto the ℓ_1 ball is a method used to enforce sparsity in a vector by projecting the vector onto the ℓ_1 ball. Detailedly, in our implementation, we give a function which takes the design matrix, outcome variable, penalty, radius of the ball, stepwise parameter, and missing rate as input parameters, and returns the estimated coefficients as output.

6. Can you provide more details on how the voting regressor and meta-analysis were incorporated into the prediction model framework?

To improve the accuracy of car price predictions, we incorporated the voting regressor, which is an ensemble learning technique that combines the predictions of multiple individual regression models to get a final prediction. In our design, we wish to select the data imputation methods and individual models for the voting regressor. In this process, we would apply several models to the complete data set obtained from data imputation and evaluate their performance using cross-validation, and compute their R-Squared scores. By selecting four models which gave the highest R-Squared scores, we may further build our voting regressor by combining these models. More specifically, in this used-car dataset, the combination of Lasso with missing data, K-Neighbors Regressor, Lasso Regression, and Linear Regression composed the final voting regressor. We experimented with this voting regressor and fitted it to the training data, evaluated its performance on the test set, and obtained an R-Squared score of 68.00%. To further enhance the model's performance, we used meta-analysis, a statistical technique that combines the predictions of the voting regressor with those of other regression models. We used the predictions generated by the voting regressor as additional features and trained a random forest regressor on the resulting dataset. By leveraging the strengths of both linear and nonlinear models, we could achieve a more comprehensive and accurate prediction of certain features. In this used-car dataset setting, we finally created a non-linear regression meta-model object and used the trained meta-model to make predictions on the test data, which had an R-Squared of 70.80%. By combining the voting regressor and meta-analysis, the final prediction model gained an accuracy of modeling fitting from 66.68% to 70.80%.

7. How did you measure the performance of your proposed model and how does it compare to other existing models?

We use R-Squared to measure the model performance and evaluate the performances of basic models before applying voting regressor based on cross validation. R-Squared is a statistical measure that represents the proportion of variance in the response variable that can be explained by the predictor variables and implies how well the model fits the data. It ranges from 0 to 1, and higher value is better since more variance in the response variable can be explained. Compared with basic existing models, the improved Lasso with missing data model achieves higher R-Squared score. The performance is further improved after utilizing voting regressor and meta analysis. The improvement verifies that our proposed model outperforms the common basic models.

8. Can your model be applied to other regions or datasets outside of North Carolina?

Yes, our statistical model can be applied to other regions or datasets outside of North Carolina. However, it is important to note that used car prices may vary by region, so some adjustments or Fine-tuning may be necessary for accurate predictions in different regions or different data sets. Additionally, the most important thing about our model is not just its ability to predict used car prices in North Carolina, but also its newly promoted idea, which includes automatic variable selection, a comprehensive understanding of feature relationships, and incorporation of voting regressor and meta-analysis. We can generalize these ideas to other prediction tasks and datasets, especially those with large amounts of missing data, to improve the accuracy and reliability of predictions. Although performing the model may vary depending on the specific dataset and the nature of the problem being solved, we believe our model could have outstanding performance in various data settings.

9. Are there any future improvements to your proposed model?

The current model of Lasso with missing data considers the missing rate of the whole data set. However, we notice that missing rates of different columns can vary. One possible future improvement is to incorporate missing rates of each column into the model so that we can get a better and more accurate description of missing data. The other potential improvement is that we can specify different weights when ensembling basic models in voting regressor.

10. How reliable are the predictions from your proposed model?

We can evaluate the reliability of the predictions from our proposed model through various performance metrics such as mean squared error, R-Squared, and root mean squared error. The performance and reliability of our model may vary depending on the specific dataset and the nature of the problem being solved.

In our second-hand car price prediction case, we have used R-Squared as the performance metric to evaluate the reliability of our model. Our proposed model, which includes an improved Lasso technique within a missing data framework and a voting regressor with meta-analysis, achieved an R-Squared score of 70.80% on the test data. This indicates that our model can explain approximately 70.80% of the variability in used car prices in North Carolina.

While an R-Squared score of 70.80% is considered to be a good fit for a prediction model, it is important to note that the performance of the model may vary depending on the specific dataset and the nature of the problem being solved. Therefore, it is recommended to further validate the reliability of the model on new and unseen data before making any critical decisions based on the model's predictions.

11. What are the practical implications of your proposed algorithm and how can it be used in real-world scenarios?

Our proposed algorithm has wide applications where there exist needs to deal with missing data. One specific aspect where this algorithm can be particularly useful is used car price prediction. When a used car is put up for sale, it may not include all the necessary characteristics required for a complete analysis, while we still want to make use of data rows with missing data and get a more precise price prediction. Under this circumstance, our algorithm can be used to incorporate missing data into regression model fit and combining results from several models. The algorithm can handle various types of missing data patterns and is more reliable than deletion or mean imputation. Moreover, the algorithm is not limited to car price prediction and can be applied to other fields where dealing with missing data is a challenge. This includes fields such as medical research, social science, finance, and marketing.

12. Are there any next steps or future research directions for this analysis?

In terms of future research directions, there are several potential steps to further develop our analysis. One crucial direction is to validate our proposed model on larger and more diverse datasets to test its generalizability to other prediction tasks where missing data is prevalent, such as in healthcare or finances, while for now, we only tested our proposed method in the used car price prediction dataset solely. Another direction is to create a user-friendly package that automates the data imputation, voting regressor, and meta-analysis, making it more accessible to a wider audience. To achieve this, we plan to develop an algorithm that can test the performances of different combinations of data imputation methods and regression algorithms including our

self-created Lasso-missing method, select the best-performing algorithms, and build the voting regressor and meta-analysis automatically. By conducting these future research steps, our model can be more widely applied and facilitate prediction tasks with missing data.

Technical appendix

Data Description and Pre-processing

The used cars dataset from Kaggle provides valuable information of second-hand car prices sourced from Craigslist, one of the largest websites for used vehicles for sale. The raw data set has 426880 rows and 26 columns. The price column serves as the outcome variable, while the remaining 25 columns are raw features, with the majority of them being categorical variables. This dataset is a valuable resource for exploring and analyzing factors that influence the pricing of used cars in the US market.

We began by filtering the data to only include entries from North Carolina, and then proceeded to drop columns that are unlikely to have a significant impact on the prediction of car prices. Specifically, we removed the id, url, description, VIN, longitude, latitude, posting date, state, and region columns. To further simplify the prediction model, we removed the model column and decided to retain only the manufacturer column, since it is more representative of the car. In total, the resulting dataset contained 13 features, which are price (the dependent variable), year, manufacturer, condition, cylinders, fuel, odometer, title status, transmission, drive, size, type, and paint color.

Through exploratory data analysis and visualization of the distribution of car prices, we observed that the price column contained a large number of outliers. These outliers would have a significant impact on the prediction if they were not processed. Upon further investigation, we noticed that there were many 0 and extremely high values in the price column, which might have resulted from errors during data scraping or issues with the data source.

To handle the outliers in the price column, we removed any prices greater than 1 million dollars. Next, we identified the remaining outliers by checking for values outside the interval defined by the median plus or minus the median absolute deviation. This approach allowed us to detect outliers that deviated significantly from the norm. Ultimately, we found that the remaining prices ranged from 1950 dollars to 33000 dollars, which fell within a reasonable range of used car prices.

To handle missing data, we dropped any rows with missing values in the year, manufacturer, fuel, odometer, title status, and transmission columns, as these features had a missing rate of less than 4%. However, even after this step, the processed dataset still contained a large number of missing values. Table 1 provides a summary of the percentage of missing values for the remaining features, which are condition, cylinders, drive, size, type, and paint color.

Table 1: Summary of missing rate.

feature	missing rate (%)
condition	38.8
cylinders	35.4
drive	33.1
size	64.1
type	22.2
paint color	26.4

Table 1 shows that the missing rate for the remaining features, namely condition, cylinders, drive, size, type, and paint color, ranges from 20% to 65%. Given the relatively high proportion

of missing values, it may not be appropriate to simply remove all the missing data in these features as this could result in valuable information being lost. Data imputation is one possible approach to handling missing data[1]. However, since these features are categorical variables and the missing rates are high, imputation methods could significantly affect the prediction if they are not carefully chosen.

Here we also present the visualizations for EDA.

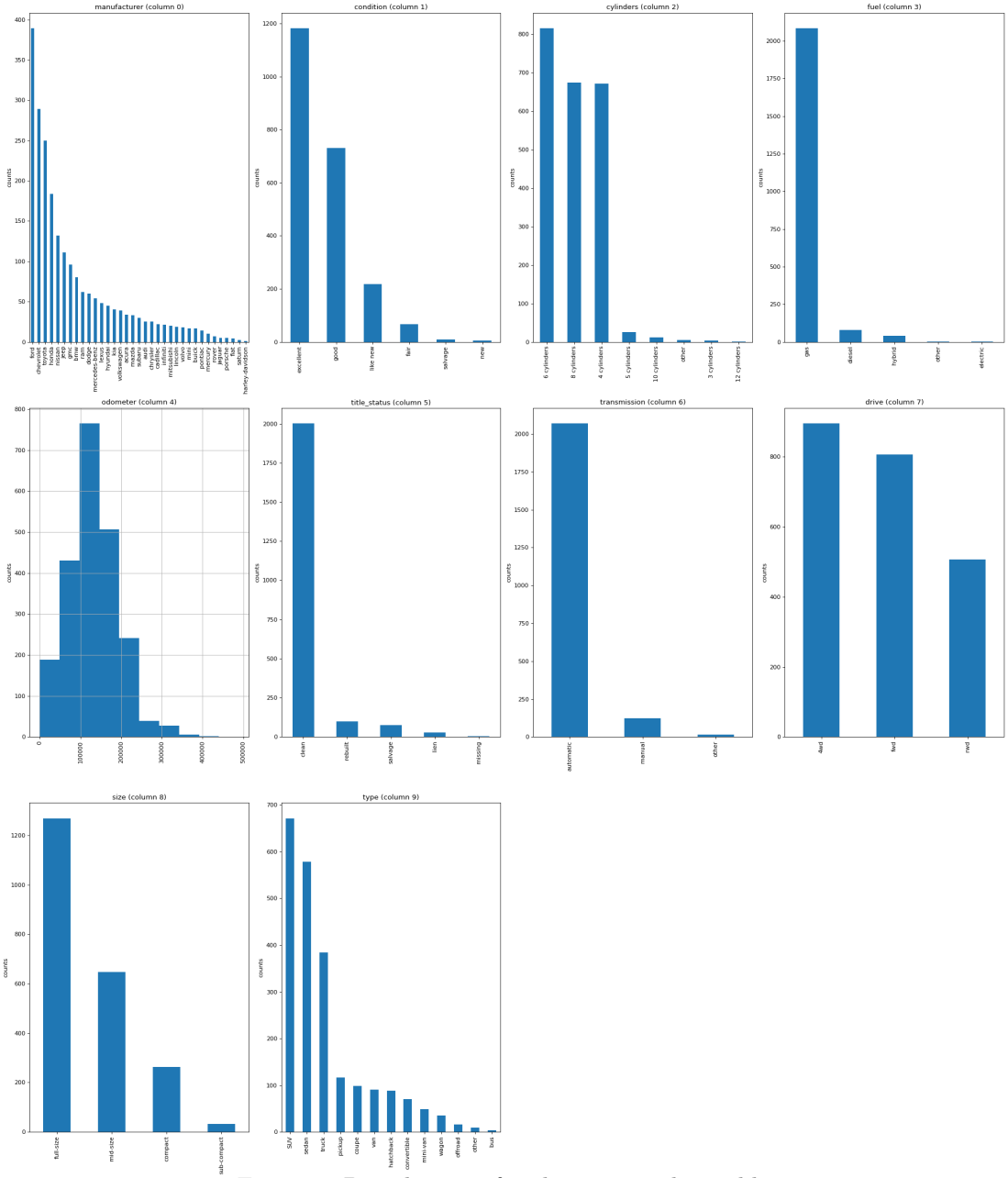
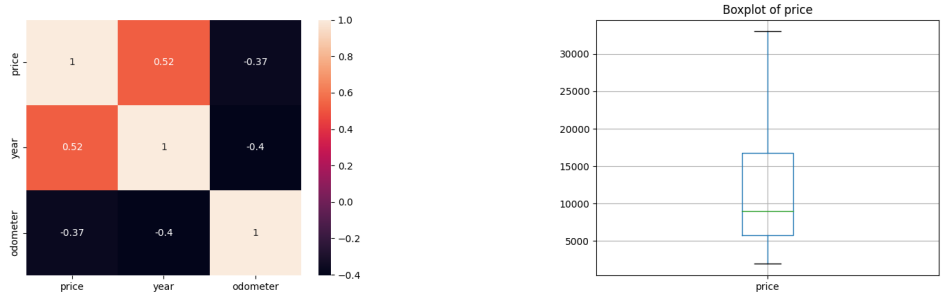
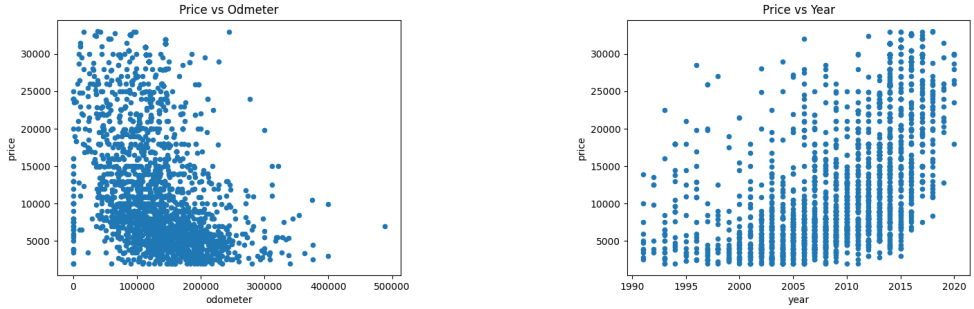


Figure 1: Distribution of each categorical variable



(a) Correlation between numerical variables (b) Boxplot of price

Figure 2: Distribution of price and its correlation with other variables



(a) (b)

Figure 3: Distribution of price vs year and odemeter

Overview

Given the nature of the data and the specific context of the problem at hand, we provided a statistical model that performs prediction under the framework of missing data problem. We proposed an improved Lasso which treats missing values as a component in model design, built the algorithm by hand in Python, and applied it to the used cars data. In addition to Lasso with missing data, we also applied and compared across basic regression models after imputing the missing data. It turns out that Lasso with missing data outperforms the basic regression models after data imputation. Based on the performances on Lasso with missing data as well as basic regressions, we further implemented a voting regressor which is as an ensemble meta-estimator. To further improve the prediction, we integrated meta analysis into our framework of prediction. Our statistical model offers several advantages: (1) Firstly, we implemented an improved Lasso technique within a missing data framework, which enables automatic variable selection. As a result, we can identify the key variables that impact car prices based on the interpretable prediction model. (2) Secondly, our approach provides a more comprehensive understanding of the relationships between the available features and car prices, taking into account the missing data in the dataset. This improves the accuracy and reliability of our predictions for second-hand car prices in North Carolina. (3) Thirdly, we incorporated voting regressor and meta-analysis into our prediction model framework, which further enhanced the prediction. (4) Finally, our model can be generalized to other prediction tasks and is particularly useful when dealing with large amounts of missing data.

The following sections describe our proposed statistical model in detail, including the relevant theoretical framework and proof. We then apply this model to the second-hand cars dataset, using the available features to make predictions about the prices of these vehicles and present the results after ensembling methods using voting and meta analysis.

Model

The real-world data set usually involves a large number of missing data, which can bring a great challenge to model prediction. As for the regression model, the standard Lasso regression is performed on the complete data set, which automatically conducts variable selection under high dimension regression by adding a penalty to the minimizing function and shrinking some of the coefficients to 0. However, the used car data set posed two challenges for model design and algorithm implementation: (1) firstly, the data set had a high number of missing values, particularly for important features such as car size and condition. (2) secondly, using Lasso regression within a missing data framework may not always result in a convex optimization problem. To address these issues, we opted to incorporate missingness into model design when improving the Lasso regression, which avoids introducing bias resulted from data imputation or discarding missing data. Additionally, we also addressed the non-convexity issue in our approach and further details are included in the following section.

Problem setup

Suppose the design matrix X is a $n \times p$ matrix and the outcome variable $y_i \in \mathbb{R}$. When the majority of features are categorical variables with multiple categories, p can be large after dummy variables are created. We first define matrix Z with entries

$$Z_{ij} = \begin{cases} X_{ij}, & \text{if } X_{ij}\text{is observed} \\ 0, & \text{otherwise} \end{cases}$$

By adding a ℓ_1 penalty and incorporating missing data into the model design, the minimizing function is formulated as:

$$\min_{\beta \in \mathbb{R}^p} \left(\frac{1}{2} \beta^T \tilde{\Sigma} \beta - \tilde{\alpha}^T y + n\lambda \|\beta\|_1 \right) \tag{1}$$

where

$$\tilde{Z} = \frac{Z}{1-\rho} \quad (\rho \text{ is a parameter related to missing rate})$$

and

$$\tilde{\Sigma} = \frac{\tilde{Z}^T \tilde{Z}}{n} - \rho \cdot \operatorname{diag}\left(\frac{\tilde{Z}^T \tilde{Z}}{n} \right), \quad \tilde{\alpha} = \frac{\tilde{Z}^T y}{n}.$$

As can be seen from the formulation, when $\rho = 0$ (that is, there is no missing data), the model degenerates to the standard Lasso problem. Hence, the improved Lasso under missing data framework can be regarded as a generalization of standard Lasso.

As pointed out in [7], the minimization problem (1) may not be a convex problem when the missing rate in design matrix X is high. Specifically, when there are a large number of missing data, the defined matrix $\tilde{\Sigma}$ may not be positive definite, leading to the presence of negative eigenvalues. To solve for the non-convexity issue, we combine the problem (1) with the algorithm proposed by [5] and make use of a composite version of projected gradient descent algorithm to deal with the minimization given by (1).

Algorithm: Lasso with missing data

According to [5], the $(t+1)$th update for the minimization problem (1) is:

$$\beta^{t+1} = \operatorname{argmin}_{\|\beta\|_1 \le R} \left(\mathscr{L}_n(\beta^t) + \langle \nabla \mathscr{L}_n(\beta^t), \beta - \beta^t \rangle + \frac{g}{2} \|\beta - \beta^t\|_2^2 + \lambda_n \|\beta\|_1 \right) \tag{2}$$

where $g > 0$ is the stepwise parameter in gradient descent algorithm, R is the radius of ℓ_1 ball, and $\mathscr{L}$ is the quadratic loss function with the gradient $\nabla \mathscr{L}_n(\beta^t) = \tilde{\Sigma}\beta^t - \tilde{\alpha}$.

By noticing that

$$\langle \beta^t, \nabla \mathscr{L}_n(\beta^t) \rangle = \nabla_\beta \mathscr{L}_n(\beta^t) = \mathscr{L}_n(\beta^t) + \langle \nabla \mathscr{L}_n(\beta^t), \beta - \beta^t \rangle$$

the update (2) is equivalent to the following update:

$$\beta^{t+1} = \text{argmin}_{\|\beta\|_1 \leq R} \left(\langle \beta^t, \nabla \mathscr{L}_n(\beta^t) \rangle + \frac{g}{2} \|\beta - \beta^t\|_2^2 + \lambda_n \|\beta\|_1 \right) \tag{3}$$

The formulation given by (3) is a composite version of projected gradient algorithm. As proposed in [2] and [3], a simple version of (3) makes use of Euclidean projection to update β^t. Since the difference between the simple version and our composite version of minimization is the addition of regularization term, we perform the update in (3) by combining soft thresholding with Euclidean projection onto the ℓ_1 ball. The implementation details and mathematical supports are included below.

Lemma. The Euclidean projection onto the ℓ_1 ball can be characterized in terms of soft thresholding.

Proof. Let Π denote the Euclidean projection and $C = \{x \in \mathbb{R}^n | \|x\|_1 \leq 1\}$ denote the unit ℓ_1 ball. By definition, the Euclidean projection of $y \in \mathbb{R}^n$ onto the set C is

$$[\Pi(y)] = \text{argmin}_{x \in C} \|x - y\|_2 = \text{argmin}_{\|x\|_1 \leq 1} \frac{1}{2} \|x - y\|_2^2$$

Based on the Lagrange multiplier, define

$$\mathcal{L}(x, \lambda) = \frac{1}{2} \|x - y\|_2^2 + \lambda \left(\|x\|_1 - 1 \right) = \sum_{i=1}^{n} \left(\frac{1}{2} (x_i - y_i)^2 + \lambda |x_i| \right) - \lambda$$

The derivative of $\frac{1}{2}(x_i - y_i)^2 + \lambda |x_i|$ is

$$\partial(\frac{1}{2}(x_i - y_i)^2 + \lambda |x_i|) = x_i - y_i + \lambda \text{sign}(x_i)$$

Since $0 \in \partial(\frac{1}{2}(x_i - y_i)^2 + \lambda |x_i|)$, we further obtain $y_i \in x_i + \lambda \text{sign}(x_i)$. Therefore, we have

$$x_i = \text{sign}(y_i)(|y_i| - \lambda)_+$$

That is, the Euclidean projection is in essence the soft thresholding. □

Based on the Lemma, we now give the implementation detail of projected gradient descent algorithm in each iteration. The update (3) is decomposed into two steps:

- (a) Soft threshold the vector $\beta^t - \frac{1}{g} \nabla \mathscr{L}_n(\beta^t)$ at level λ_n: $v = \textit{soft-threshold}(\beta^t - \frac{1}{g} \nabla \mathscr{L}_n(\beta^t), \lambda_n)$
- (b) Project the obtained vector onto the ℓ_1 ball to obtain $v' = \Pi(v)$ if $\|v\|_1 > R$.

With algorithm proposed as above, we built the algorithm in Python and applied it to the used car data set. The main function in our algorithm takes design matrix X, outcome variable y, penalty l, radius of the ℓ_1 ball R, stepwise parameter g, missing rate ρ as input parameters, and returns the estimated coefficients β as output. The two steps in each iteration are achieved based on two helper functions, one implements soft threshold and the other corresponds to Euclidean projection onto ℓ_1 ball. Based on exploratory data analysis and with dummy variables created,

the design matrix X has 9274 rows and 86 columns. Here, we set the input ρ to be 0.12 which is the average missing rate of matrix X. One future improvement of our model is to incorporate the missing rate of each column into the model. We then applied the algorithm to the data set to estimate the coefficients β and selected R^2 as the measure of goodness-of-fit for the model. The performance of Lasso with missing data together with other basic regression models are summarized in Table 2.

Voting and Meta Analysis

To improve the accuracy of our car price prediction, we decided to use a more advanced approach based on Lasso with missing data as our final model. We chose to first implement a voting regressor, which is an ensemble learning technique that combines the predictions of multiple individual regression models to obtain a final prediction.

To select the individual models for our voting regressor, we applied several models to the complete data set obtained from data imputation. Specifically, we applied k-Nearest Neighbor (KNN) Imputation [6] and Hot Deck imputation [4]to impute the missing data in the cars data set. KNN imputation works by using the k-nearest neighbors algorithm to identify the most similar (nearest) observations to the one with missing data. The missing value is replaced with the average (for numerical variables) or mode (for categorical variables) of the identified neighbors. Hot Deck imputation works by fill in missing entries with a value randomly selected from the corresponding variable. We then evaluated the performances of models using cross-validation and computed their R-squared scores. The results are summarized in Table 2.

Table 2: Summary of basic models

Models	R^2 (%)
Linear Regression	65.87
KNeighborsRegressor	59.27
Lasso with Hot Deck Imputation	60.63
Ridge	65.42
Bayesian Ridge	65.44
ElasticNet	61.57
Lasso	65.44
Lasso+missing	**66.68**

Based on these results, we selected four models which give highest R-squared score when used in the Voting Regressor. It turns out to be the combination of Lasso with missing data, KNeighborsRegressor, Lasso Regression and linear regression.

Figure 4 shows the results of using four different regressors to make the first 20 predictions on the test data. The two versions of Lasso regression (standard Lasso on complete data set obtained from data imputation and Improved Lasso with missing data) as well as linear regressor produce very similar predictions while the predictions of KNeighbors Regressor turn out to be more unstable, with larger variations in predicted values for test samples. Despite the differences in stability, all regressors appear to perform reasonably well, with predicted values generally close to the true values for each test sample. It is worth noticing that the performance of the regressors may vary depending on the specific dataset and the nature of the problem being solved when generalizing to other prediction tasks with other data sets.

In addition to these individual regressors, we also experimented with a voting regressor that combines the predictions of multiple regressors. This approach can potentially improve the overall performance by taking advantage of the strengths of each individual regressor.

After fitting the voting regressor to the training data, we evaluated its performance on the test set and obtained an R-squared score of **68.00%**. This result indicates that the voting regressor

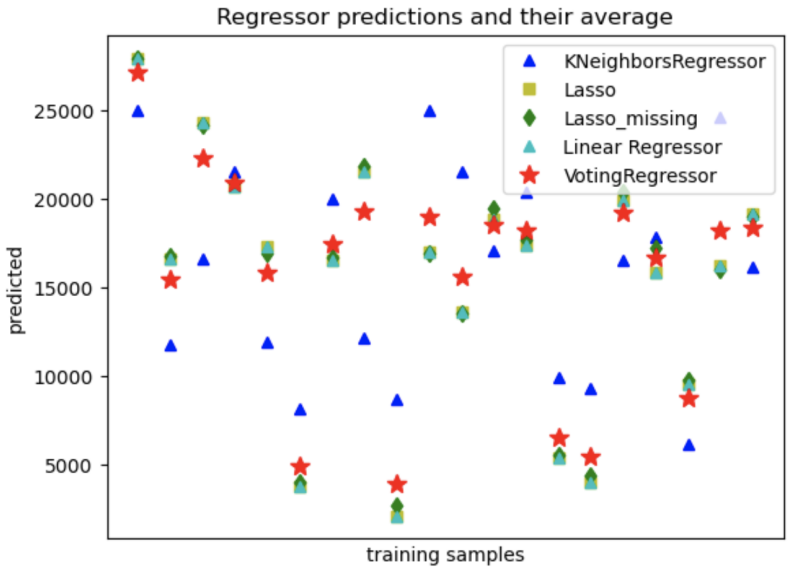


Figure 4: Models in Voting Regressor

was able to capture the predictive power of the selected linear regression models and generated more accurate predictions compared to any individual model.

To further enhance the model performance, we utilized meta-analysis, a statistical technique that enables us to combine the predictions of our voting regressor with those of other regression models. To achieve this, we used the predictions generated by our voting regressor as additional features and trained an extra trees regressor (random forest regressor) on the resulting dataset. This approach enabled us to capture any nonlinear relationships that may have been overlooked by our linear models and improve the accuracy of our final prediction. By leveraging the strengths of both linear and nonlinear models, we were able to obtain better model performance on used car data.

We created a non-linear regression meta-model object and used the trained meta-model to make predictions on the test data. The R-squared of the meta-model on the test data was calculated to be **70.80%**, indicating that the meta-model has a good fit with the data.

By combining with voting regressor and meta analysis, our final prediction model gains the accuracy of modeling fitting from 66.68% to 70.80%. Table 3 summarizes the performances in terms of R-squared of different models. Overall, using a voting regressor and a meta-model improves the performance of car price predictions by combining the strengths of multiple individual regression models after implementing the improved Lasso with missing data.

Table 3: Summary of results.

Models	R^2 (%)
Lasso + missing	66.68
Voting (combination of the three models)	68.00
Voting + Meta Analysis	70.80

References

[1] M. Abonazel. Handling outliers and missing data in regression models using r: Simulation examples. *Academic Journal of Applied Mathematical Sciences*, 6:187–203, 09 2020.

[2] A. Agarwal, S. N. Negahban, and M. J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery, 2012.

[3] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the l1-ball for learning in high dimensions. In *ICML '08: Proceedings of the 25th international conference on Machine learning*, pages 272–279, New York, NY, USA, 2008.

[4] D. Joenssen and U. Bankhofer. *Hot Deck Methods for Imputing Missing Data*, page 63–75. 07 2012.

[5] P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3), jun 2012.

[6] D. M. P. Murti, U. Pujianto, A. P. Wibawa, and M. I. Akbar. K-nearest neighbor (k-nn) based missing data imputation. In *2019 5th International Conference on Science in Information Technology (ICSITech)*, pages 83–88, Yogyakarta, Indonesia, 2019. IEEE.

[7] R. J. Tibshirani and L. A. Wasserman. Sparsity, the lasso , and friends statistical machine learning, spring 2017. 2017.