Case 1

02582 Computational Data Analysis

s230284 & s233091 (Group 5)

March 17, 2024

Introduction/Data description

In the given statistical learning problem, our objective is to predict the target variable y based on $N = 100$ observations, each with $p = 100$ features. This scenario, where $N \approx p$, typically results in a high-dimensional data space that can lead to high variance and overfitting issues. High variance in this context implies that our model might perform well on the training data but poorly on unseen data, due to its excessive complexity and potential capture of noise as signals.

Before selecting an appropriate model, it is crucial to perform some initial exploratory data analysis. This includes examining correlations among features to identify potential multicollinearity, which can affect model performance by making the estimation of coefficients unstable. Assessing the number of missing values is essential for determining the need for imputation techniques or the removal of certain observations or features. Finally, locate and handle categorical features.

Continuous

Missing values & Preprocessing

In this section we will examine the presence and frequency of the missing values within our dataset for the continuous features. In Fig. 2 and Fig. 1, we see that the number of missing values for both rows and features do not exceed 25 which is a moderate amount of missing values compared to the size of the dataset.

For fitting the model later, we decided to drop the features with missing values with zscore > 2 in attempt to reduce the dimensionality while also handling partially some of the missing values. The results are visualized in Fig 3.

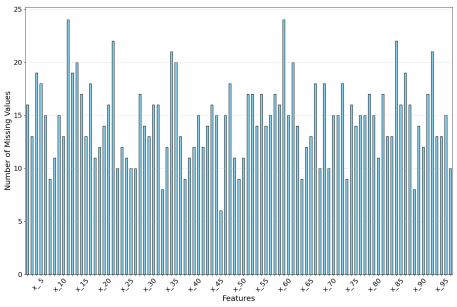

Figure 1: Missing values per feature

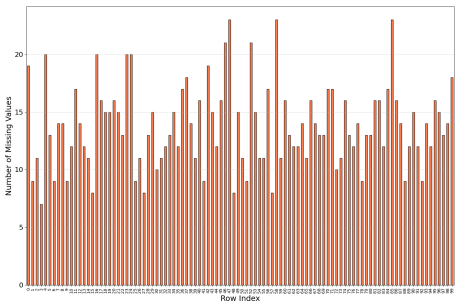

Figure 2: Missing values per row

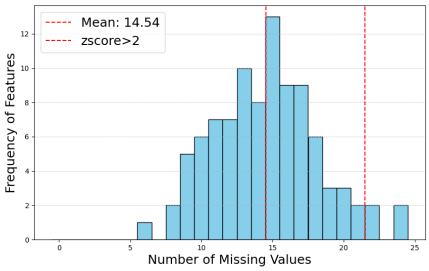

Figure 3: Histogram of missing values

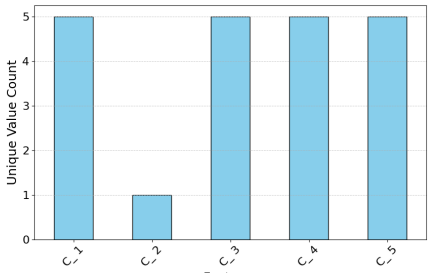

Figure 4: Unique value count

Categorical

Missing Values & Preprocessing

In this section we examine the presence of categorical features in our dataset. There are 5 categorical features with their unique value counts displayed in 4. By observing Fig3, we see that C1 has a lot of missing values so we decided to drop this feature. Also, C2 has only one unique value (see Fig4), so we emit using it in the fitting process. We also performed one hot encoding at.

Model

Given the scenario where we have a limited number of observations compared to the number of features, a situation further worsened by one-hot encoding of categorical variables, it is crucial to choose a model that minimizes the risk of overfitting while maintaining the ability to generalize well to new data. Due to its simplicity, one might initially consider an OLS Linear Regression model to be appropriate. However, the presence of multicollinearity, as indicated by the significant correlation among features in our dataset (see Fig. 6,5), makes necessary the use of some kind of constraint for the coefficients. Multicollinearity

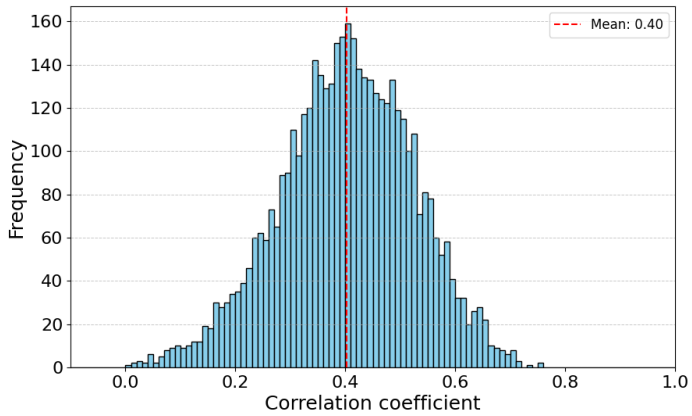

Figure 5: Correlation histogram

can lead to inflated variances for the estimated coefficients, making them highly sensitive to changes in the model.

To address these concerns, we decided to use Elastic Net Regularization along with the Coordinate Descent Algorithm, a hybrid approach that combines the strengths of Lasso (L1 regularization) and Ridge (L2 regularization) penalties. This method allows for both variable selection and shrinkage, enabling us to deal with the issues stemming from multicollinearity and high dimensionality. By using Elastic Net, we aim to achieve a compromise between bias and variance, selectively shrinking some coefficients towards zero (like Lasso) while also distributing the penalty across all coefficients (like Ridge).

To optimally select the regularization parameter λ and the mixing parameter α, which balances the Lasso and Ridge penalties in the Elastic Net model, we employed a two-step grid search strategy. Initially, a broad search was conducted over 200 values each for λ (ranging from 0 to 100) and α (ranging from 0 to 1) to identify a region of interest in the parameter space. Subsequently, a finer grid search was performed around the best parameters identified in the initial step to precisely determine the optimal values.

Prior to fitting the model, the data were standardized to have mean 0 and variance 1. The dataset was then partitioned into training and testing sets, with 20% of the data reserved for testing. The model fitting was conducted on the training set testing both 5-fold cross-validation to assess the average error associated with each parameter combination.

During the initial broad grid search, the optimal parameter combination was identified as $\lambda = 1$ and $\alpha = 0.8$. This guided the subsequent, more focused search, where the parameter space was narrowed to λ values between 0.1 and 2 (200 values) and α values between 0.7 and 0.9 (20 values).

The refined search resulted in an optimal parameter set of $\lambda = 0.6$ and

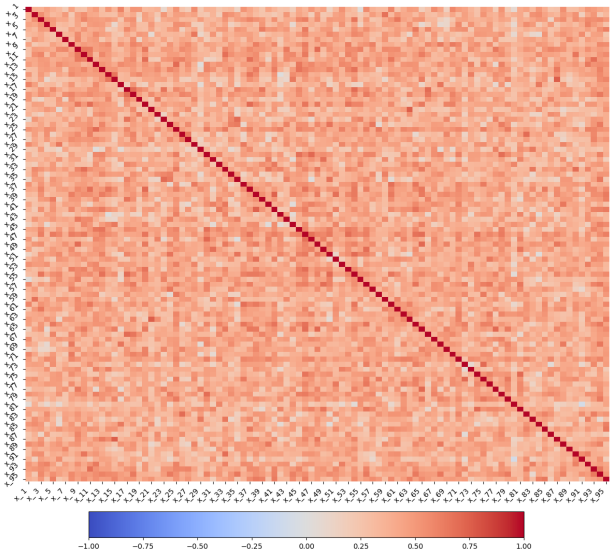

Figure 6: Correlation heatmap

$\alpha = 0.9$. This combination achieved a mean squared error (MSE) of 514.5 on the test set. Importantly, this parameter configuration led to a model where 46 features were effectively reduced to zero. The resulting top 10 model coefficients are illustrated in Fig. 7.

Estimated RMSE

We are going to employ a bootstrap technique to estimate the RMSE of the new data or the equivalent EPE. This approach involves generating samples that are equal in size to our original dataset(100), thereby simulating new data instances from the existing dataset with known targets. For each bootstrap sample, we perform a split into training and testing sets, fitting the model on the training set and calculating the RMSE on the test set(20%). This process was repeated for 10,000 bootstrap samples, allowing us to approximate the distribution of RMSE values we might expect when predicting on new data.

The distribution of the calculated RMSEs from these simulations is represented in Fig. 8. The mean of this distribution serves as our best estimate for the RMSE that the model might exhibit on actual unseen data, which was 21.55. This technique, indirect, provides insight into the model's potential performance in a real-world scenario where targets are not available for new data.

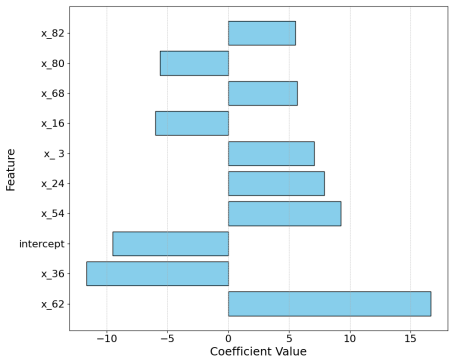

Figure 7: Model Top 10 Coefficients

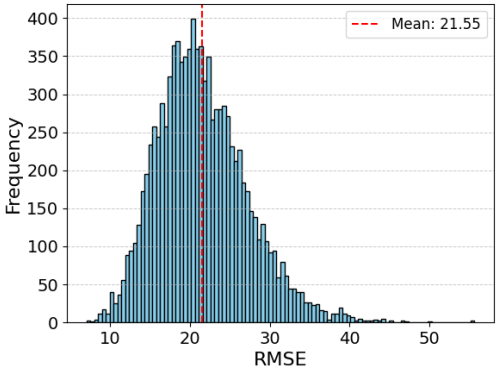

Figure 8: Histogram of RMSE values from bootstrap simulations.

Conclusion

Finally, the chosen model of Elastic Net Regularization combined with the Coordinate Descent Algorithm with optimal parameters of $\lambda = 0.6$ and $\alpha = 0.9$ was chosen, achieving a mean squared error (MSE) of 514.5 on the test set and reducing 46 features to zero.

The bootstrap technique estimated the RMSE for new data from 10,000 samples of the original dataset, as the mean of the resulted RMSE distribution over the samples, which was 21.55.

For the sake of reproducibility, we provide our code [1].

[1] https://github.com/Mirtia/02582-Computational-Data-Analysis/tree/main