Case 1

02582 Computational Data Analysis

s164521 & s194654

March 19, 2024

1 Introduction and Data description

This case study aims to build a predictive model of Y. The model is build on 100 observations containing a response vector Y and a 100-dimensional feature vector X. Within the feature vector X, 95 dimensions are continuous, while the remaining 5 dimensions are categorical. The model build on the previously mentioned data, is then used to predict y-values for a 1000 new observations, represented by the same 100 features.

During the initial exploratory phase of the data, it is discovered that all continuous features, denoted as x_1 through x_{95}, exhibit instances of missing values, totaling 1381 missing values across the continuous features. Conversely, within the categorical features, only the categorical feature C_1 manifests instances of missing values, precisely amounting to 28 occurrences. When no further detail about the dataset and its features are given, it can be difficult to determine whether data are missing at random or not. Therefore, when nothing more is known, it is assumed that missing values are missing at random. When looking at the new data, one must be aware of missing values in more than just C_1. This must be accounted for when handling missing values and categorical features as it otherwise will impact the generalization of the model, as the number of categories within the different categorical features will be different from for the two different datasets.

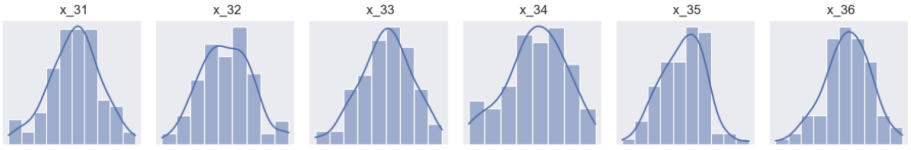

Figure 1: Subset of features and their distribution

From Figure 1 we see a subset of the data and their distributions, here it can be seen that the continuous data are normally distributed, although some features

are slightly skewed to the left or right. The distribution of the full dataset can be seen in **Appendix 1**. Looking at the correlation between features, only a few features have a weak negative linear correlation. Most correlations are positive and have a moderate linear correlation. There are more features with absolutely no or weak linear correlation than there are features with strong linear correlation. This means that the features are relatively independent and implies less redundancy in the model.

2 Model and method

2.1 Missing values

2.1.1 Continuous features

The dataset contains a large amount of missing values considering its size. Therefore, it is extremely important that imputation is done properly, otherwise it can introduce bias into the model. Imputation is done after the dataset is split into validation and training parts, as this introduces bias if done the other way around. This also means that when performing k-fold cross validation, imputation is performed inside each cross validation fold. This is because if imputation is done before splitting the dataset into validation and training, the training set has knowledge about the validation set, as the entire dataset is then used to fill in the missing values and therefore introduce bias in the model. This is not beneficial for a general model and therefore imputation is done after the data set is split.

Before doing imputation, the data is standardized, as the chosen imputation methods uses data in other features to impute missing values. The imputation method chosen is *IterativeImputer* as we work with multivariate data. The imputation methods work by utilizing the available data in other features, in order to impute the missing values. In more details at each step a feature column is chosen as output y while the others are considered to be inputs X. Next a regressor is fitted on (X,y) for all the known values in the chosen feature y and the regressor predicts the missing values for y. This method has proven to be better than mean and median imputation. This can also be explained by the fact that in the exploratory analysis we see some of the features with many missing values have a moderate linear correlation with other features. Figures 2 and 3 below shows a subset of the features before and after imputation and shows that the chosen imputation methods captures the trend of the data.

2.1.2 Categorical features

The method chosen to convert categorical features into numerical has a direct impact on the way missing values are processed within the categorical variables, hence some reasoning for handling missing values will also be discussed in the section regarding *Factor handling*. As our knowledge about the data is so sparse,

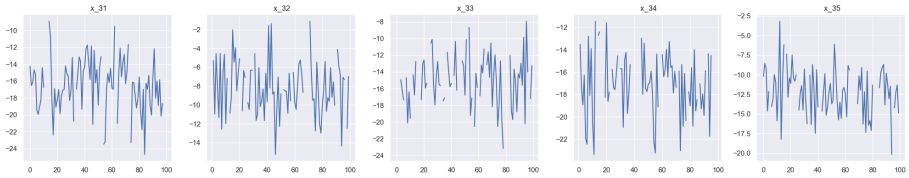

Figure 2: Numerical features with missing values

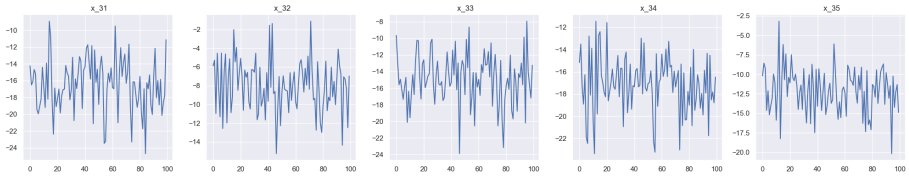

Figure 3: Numerical features after imputation

the missing categorical values are treated as its own category. This is done as it preserves the information that something is missing. When doing imputation it can introduce bias, especially if the values are not missing completely at random. As we don't know the story behind the data, we cannot conclude that for the overall dataset, hence why the missing values are treated as its own category.

2.2 Factor handling

Many machine learning models require a numerical input, hence an encoding method has been used in order to convert the categorical features into numerical ones. The chosen method to perform encoding is one-hot encoding. One-hot encoding is a binary encoding scheme, each category is represented as a binary vector with a length equal to the number of unique categories in the variable. However as we have no knowledge about the data and no knowledge about whether missing data is missing at random or not, we chose to keep the missing values as a part of the dataset. When doing so, the same vector must be introduced for all features with the same categories, in order to make the model as general as possible.

If one analyzes the data, it is expressed that in the new test data, C_2 has the value H, but is also accompanied by missing values. Based on the knowledge of the first data, you can quickly draw the conclusion that missing values in C_2 is missing completely at random, as only 'H' is represented in addition to the missing values. However, since there is no more knowledge of the data and its scope, it is decided to keep the missing values. Therefore, one-hot encoding is performed so that all categories (G, H, I, J, K and NaN) are present for all features. Using this introduces more complexity than necessary. However, it is not considered to have an impact on the bias-variance trade-off, since $p > n$ and not $p >> n$. Adding the extra dimensions should not have an effect on the

model efficiency.

2.3 Model selection

In order to chose a proper model for predicting the target variable y, a correlation analysis was carried out in order to understand how different features can effect y. Figure 7 shows that some features are significantly more correlated with y then others, where Figure 8 shows that among the top features correlated with y, some of these are also highly correlated with each other.

Based on the correlation analysis, while considering that $p > n$, a model with feature selection was chosen in order to prevent the model from overfitting. Considering the linearity of the data, an Elastic Net model was chosen, as it applies both L1 and L2 regularization in order to shrink the weights of the models while also completely removing non-important features. In order to chose the optimal Elastic Net model, a hyper parameter grid search was carried out on the regularization parameters. The model that minimized the Mean Squared Error had the values $\alpha = 0.6316$ and $\lambda = 1.0$.

2.4 Model Validation

With the optimal values for α and λ, the model training was carried out together with 5-fold Cross Validation, where the training data was standardized and the missing values were imputed within each fold. The reason for taking $k = 5$ lies within the size of the dataset. The results can be seen in Table 1 below. As the known y-values varies in a range from $[-178.05 : 113.45]$ the evaluation metrics are considered to be within a reasonable framework, with no clear sign of over-fitting.

MSE	**RMSE**	**MAE**	**R^2**
186	13.65	11.25	0.921

Table 1: Elastic Net training results

As shown in Figure 4, the Elastic Net selects a total of 30 features, where the features x_{54}, x_{62} and x_{36} are the three most important features.

3 Results

The optimal Elastic Net model can then be applied on the new unseen Test Data, consisting of the same 100 features and 1000 observations. First the test data is standardized with the mean and standard deviation of the training data, while missing values are also imputed. The predicted values are shown in Figure 5 and it can be seen that it follows more or less the same summary statistics as

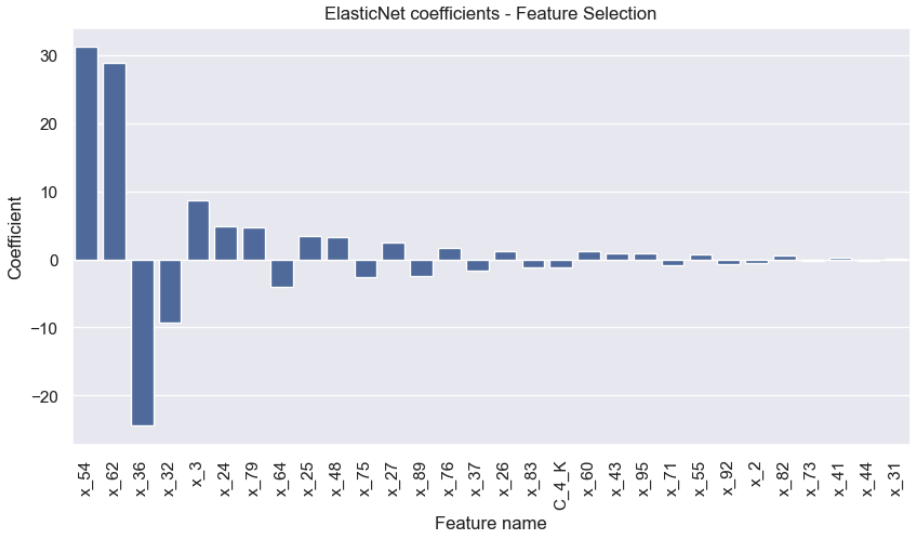

Figure 4: Elastic Net Feature Selection

the target variable of the training data set, as shown in Figure 9 in **Appendix 1**. However, in order to estimate the prediction performance, the RMSE has also been estimated.

The RMSE is estimated by considering the RMSE values obtained from each fold in k-fold cross-validation during the model training process. RMSE quantifies the average disparity between the predicted values of the model and the actual values. Therefore, by incorporating the RMSE values from all folds in k-fold cross-validation and averaging them, we arrive at an estimated:

$$RMSE = \frac{1}{5} \cdot \sum_{k=1}^{5} RMSE_k = 17.19 \tag{1}$$

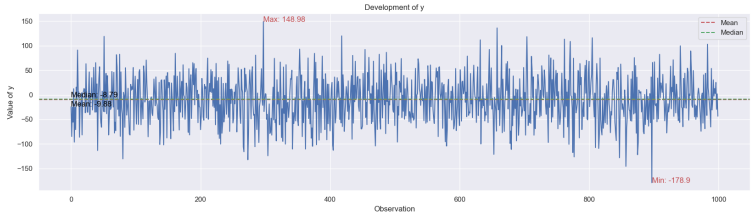

Figure 5: Predicted y-values for test data

Appendix

Appendix 1

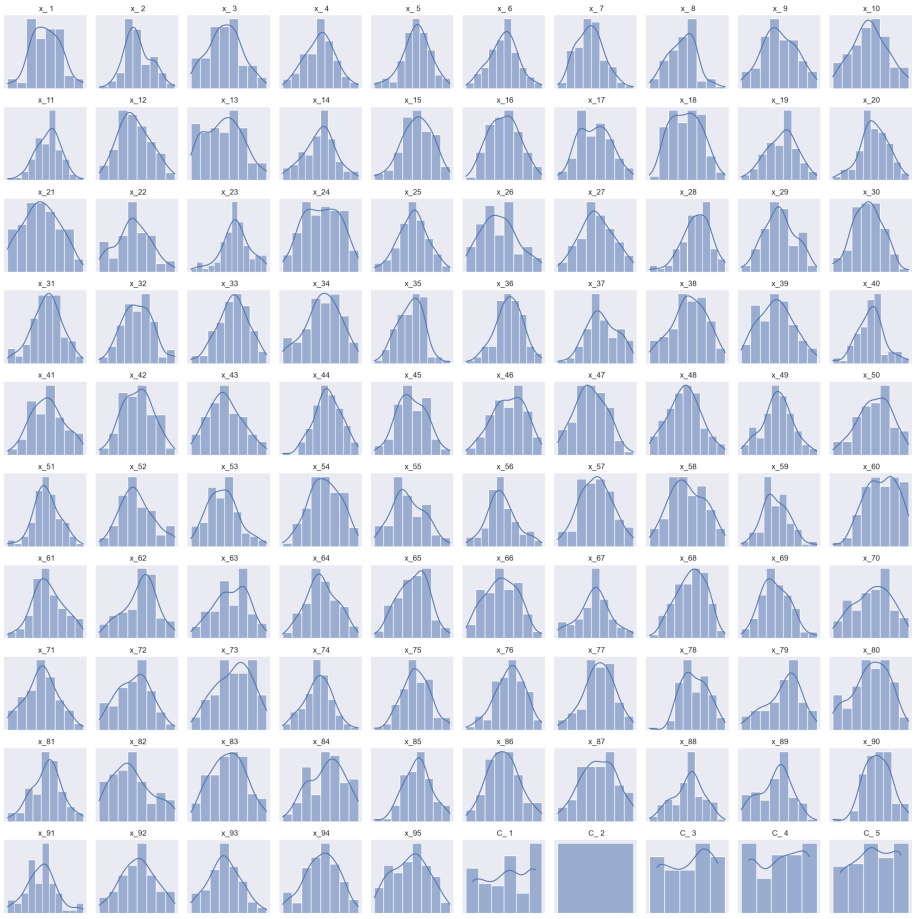

Figure 6: Features and their distribution

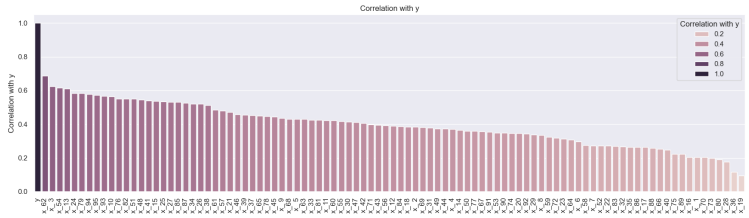

Figure 7: Training correlation with y

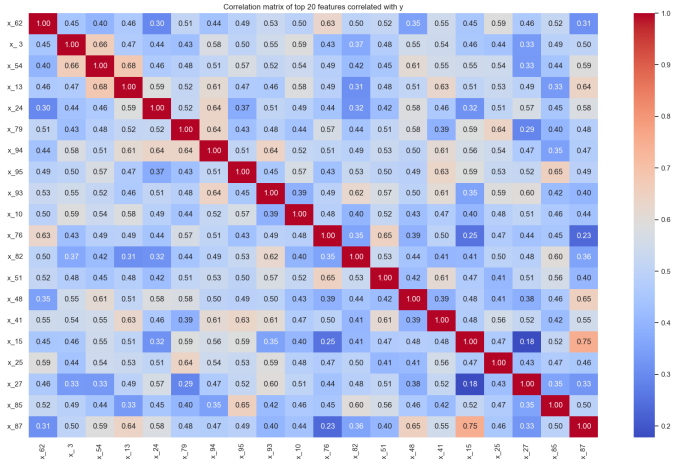

Figure 8: Training features correlation with each other

Figure 9: Development and summary statistics of y in training data