Table of Contents

1. Introduction

In the era of rapid fintech evolution, understanding and predicting ATM cash withdrawal behaviour is vital for financial institutions' strategic planning and resource optimization. This report delves into a large dataset comprising 22,000 observations, aimed at uncovering the impact of urban environmental characteristics on ATM cash withdrawal behaviours and developing accurate predictive models. The dataset includes variables directly pertinent to financial transactions, such as the number of shops, ATMs, distance to downtown, weekday traffic, central area activities, and indicators of high-value withdrawals.

Through exploratory data analysis (EDA), we initially identified the correlations and potential patterns among variables, providing intuitive understanding for subsequent model development. Further, we employed several statistical models for comparison to optimize the accuracy and interpretability of predictions. Specifically, these include:

- Ridge Regression: An enhancement to linear regression, incorporating L2 regularization to address multicollinearity.
- LASSO Regression: Utilizes L1 regulation to encourage sparse solutions, thereby performing variable selection.
- Elastic Net Regression: Combines the features of Ridge and LASSO regression using both L1 and L2 regularizations to balance feature selection with model complexity.
- OLS: Minimize the sum of the squares of the observed dependent variable in the dataset and those predicted by the linear function.

Each of these models was tuned through cross-validation and assessed for its predictive performance on unknown data. The report not only elaborates in detail the construction process, training specifics, and parameter choices of each model but also analyses their comparative performance, providing the results of the final predictive model via multi-model ensemble methods.

Through these meticulously trained models, we offer data-supported insights on how to strategically deploy ATMs in urban financial landscapes to maximize user experience and operational efficiency. The conclusions and models presented in this report serve as a reference for financial decision-makers in optimizing service layouts.

2. Exploratory Data Analysis

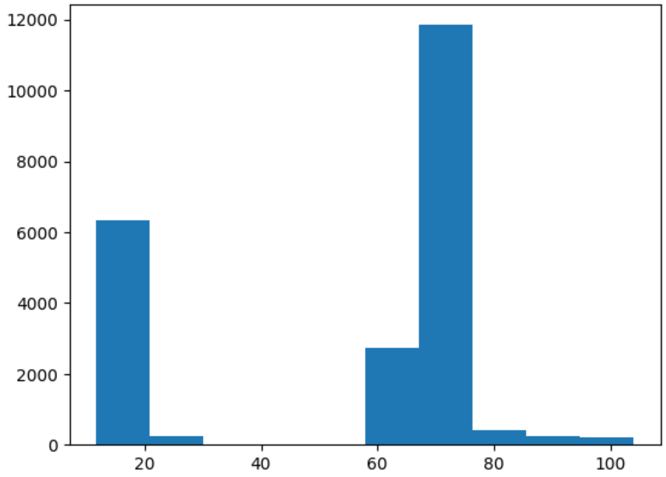


Figure 1: Histogram of spread of "Withdraw" response variable

The histogram demonstrates a notable concentration of withdrawal amounts within two distinct ranges: 8-21 and 65-75. This suggests that a considerable proportion of customers tend to make moderate-sized and relatively large withdrawals. These peaks might represent regular spending patterns, indicating common transactional behaviour among specific groups of customers.

Notably, the histogram indicates a pronounced surge in withdrawals specifically within the 68-75 range. This finding suggests a prevalent tendency among a significant portion of customers to withdraw larger sums, possibly indicating specific spending behaviours, financial commitments, or regular expenses that fall within this range.

Interestingly, there appears to be a dip in the frequency of withdrawals between the 30-58 range, indicating a relatively less common occurrence of transactions within this interval. This might signal a lull in spending activity or fewer instances of moderate withdrawals compared to other ranges.

Additionally, the presence of a small peak around 58-68 suggests a secondary pattern of moderate withdrawals. This minor peak could indicate a specific group of customers or a particular spending category that consistently withdraws moderate amounts within this range.

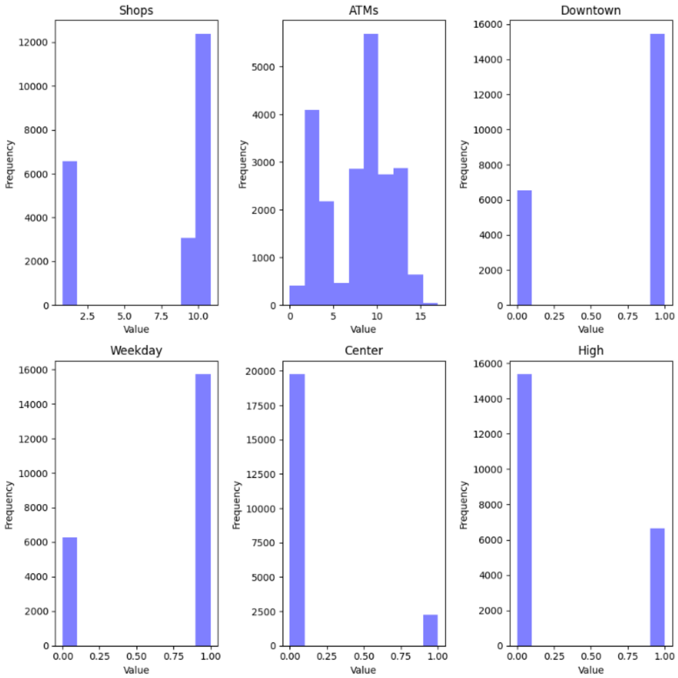


Figure 2: Histogram of spread of explanatory variables

The numerical variables "Shops" and "ATMs" indicate the availability of shops and ATMs in the area under consideration. The distribution of these variables implies that there are varying degrees of accessibility to these essential facilities. There are many Shops around 1-2, 10-11, and several shops around 9-10 with majority having 10-11 Shops. There is a varying number of ATMs with many being around 2-5 and 7-13 with majority having 10-11 ATMs.

The categorical variables "Downtown" and "Weekday" highlight the presence or absence of specific characteristics in the area. The dominance of '1' in the "Downtown" variable indicates a prevalent presence of the area being situated in or near the downtown region, potentially implying a higher level of commercial and business activities within the vicinity.

Similarly, the predominance of '1' in the "Weekday" variable suggests a higher frequency of activities or events occurring during weekdays, signifying potential regular business activities or daily routines within the area.

The categorical variables "Centre" and "High" reflect the presence or absence of specific features or characteristics. The prevalence of '0' in the "Centre" variable suggests that the area might be relatively farther from the central hub or city centre, potentially indicating a more suburban or peripheral location. Similarly, the dominance of '0' in the "High" variable implies that the area may not be situated at a higher altitude or elevation, indicating a probable lower elevation location.

2.1. Correlation between Predictors and Response Variable

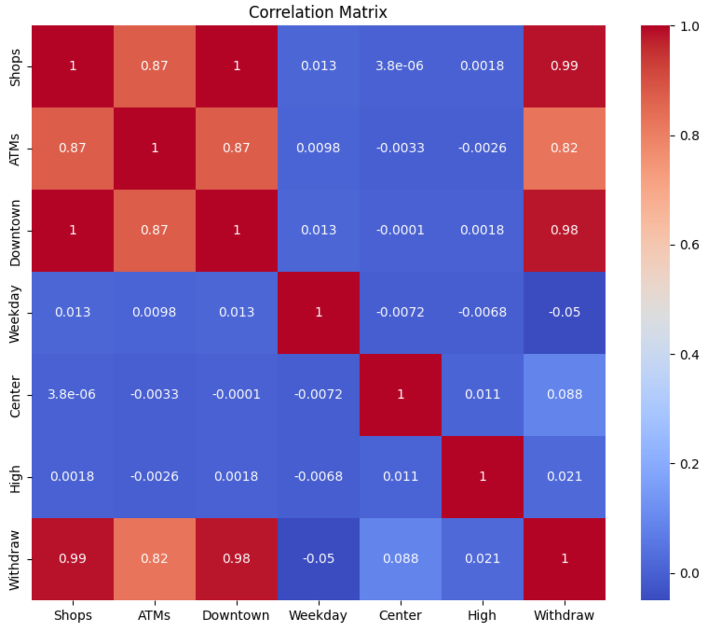


Figure 3: Correlation Matrix between all variables

Shops and Withdraw Correlation (0.99):

The high correlation between the "Shops" variable and "Withdraw" implies a strong association between the availability of shops in the area and the frequency of financial withdrawals. This suggests that areas with a greater concentration of shops are more likely to witness higher withdrawal activities, indicating a potential link between commercial activities and financial transactions.

ATMs and Withdraw Correlation (0.82):

The considerable correlation between the "ATMs" variable and "Withdraw" signifies a notable relationship between the availability of ATMs and the frequency of withdrawals. This correlation suggests that areas with higher ATM accessibility tend to experience increased financial transactions, indicating the crucial role of accessible financial services in facilitating customer transactions.

Downtown and Withdraw Correlation (0.98):

The strong correlation between the "Downtown" variable and "Withdraw" suggests a significant association between the proximity to the downtown area and the frequency of financial withdrawals. This indicates that areas in or near the downtown region tend to observe higher withdrawal activities, highlighting the influence of location on customer financial behaviours.

Multicollinearity Challenges between Predictor Variables:

The high correlations between predictor variables, such as "Shops" and "Downtown," as well as "Shops" and "ATMs," and "ATMs" and "Downtown," can lead to multicollinearity issues. These intercorrelations among predictors can introduce instability in the model, affecting the reliability and interpretability of the coefficient estimates. It is crucial to address these multicollinearity challenges to ensure the robustness and accuracy of the predictive model.

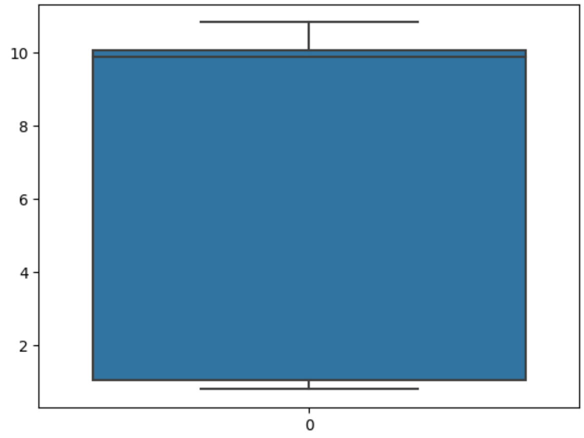


Figure 4: Box Plot of Shops

The box plot demonstrates a significant concentration of the number of shops within a relatively narrow range, with the median positioned close to the upper quartile of the data distribution. This positioning suggests that a substantial proportion of the area under study hosts a notable density of commercial establishments, indicating a vibrant and bustling business environment.

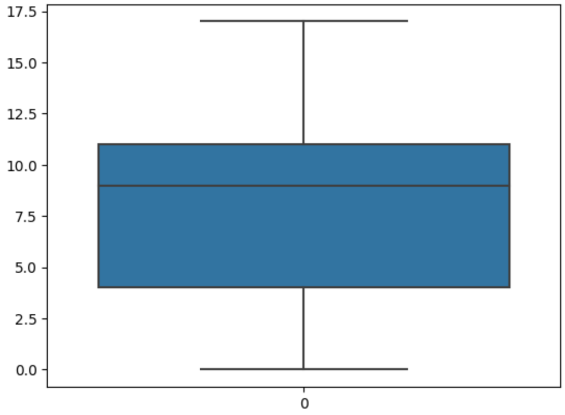


Figure 5: Box Plot of ATMs

The median value of 9 signifies the central tendency of the data distribution, suggesting that most of the area boasts a moderate accessibility to ATMs. This implies that residents and businesses in the vicinity generally have reasonable access to essential financial services, promoting convenience and facilitating smooth monetary transactions.

The extensive range spanning from 0 to approximately 17 highlights the diverse accessibility levels of ATMs across different parts of the area. This wide-ranging accessibility landscape underscores the varying degrees of financial inclusivity and infrastructure development, emphasizing the need for strategic measures to ensure equitable access to banking services for all members of the community.

The first quartile positioned around 4 and the third quartile located at approximately 11.25 illustrate the spread of ATM accessibility levels within the area. This indicates that a significant portion of the region enjoys relatively moderate to high access to ATMs, while a smaller segment may have limited access.

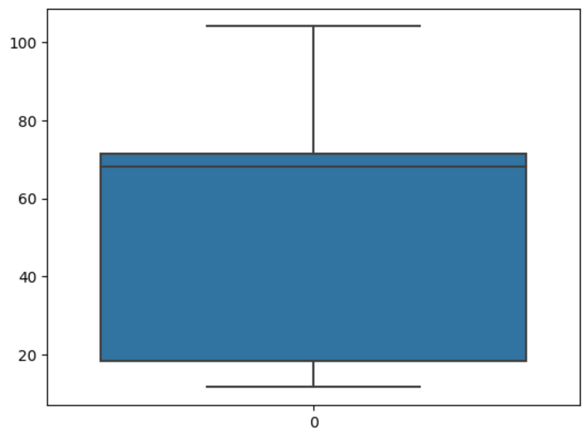


Figure 6: Box Plot of Withdraw

The proximity of the median value to the third quartile, around 70 to 73, highlights the central tendency of the withdrawal amounts. This positioning suggests that a significant portion of the observed withdrawals falls within a relatively narrow range, indicating consistent transactional behaviours or financial patterns within the dataset.

The wide-ranging values from approximately 15 to 105 signify the diverse spectrum of withdrawal amounts, reflecting the varying financial needs and preferences among the individuals or entities making the withdrawals. This wide dispersion emphasizes the importance of understanding the factors influencing these diverse withdrawal behaviours and the potential implications for financial planning and management.

The first quartile at 20 and the third quartile around 73 delineate the spread of withdrawal activities, indicating that a considerable proportion of the observed transactions fall within these ranges. This spread highlights the prevalence of specific withdrawal patterns or behaviours, underscoring the need for tailored financial services that align with the diverse transactional requirements of the customer base.

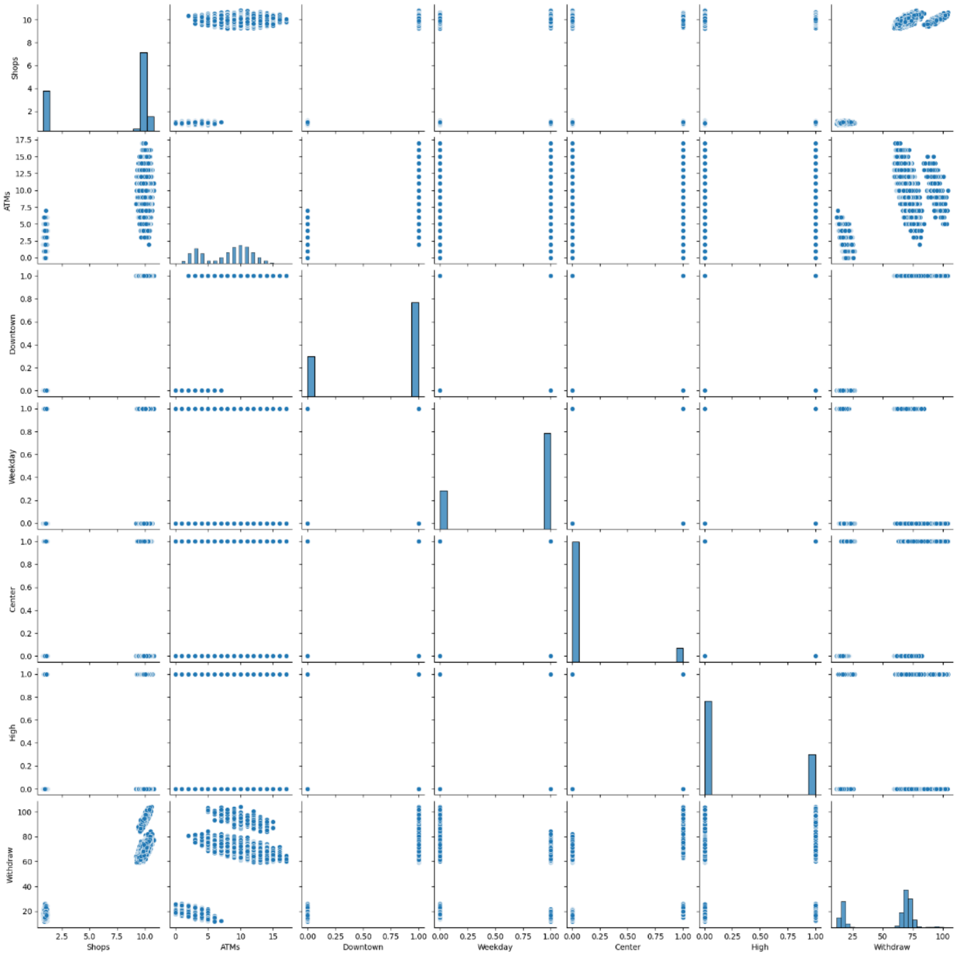


Figure 7: Pair Plot of all explanatory variables

When the number of ATMs is low (around 7.5 and below), the number of shops tends to be around 1. However, as the number of ATMs increases (between 2.5 and 17.5), the number of shops tends to cluster around 9-11. This suggests that a higher concentration of ATMs might be associated with a greater number of shops, potentially indicating a more commercial or densely populated area.

Secondly, the location of the area, whether downtown or not, plays a crucial role in the distribution of shops and ATMs. When the area is categorized as downtown (Downtown=1), both the number of shops and ATMs significantly increase. This indicates that the downtown area is a hub for commercial activities and financial services, making it a focal point for economic transactions.

Furthermore, the frequency of withdrawals appears to impact the presence of both shops and ATMs. When the number of withdrawals is low (below 30), the number of shops and ATMs is also low, around 1. However, when the frequency of withdrawals is between 50 and 110, there is a substantial increase in both shops and ATMs, clustering around 9-11.

Moreover, the day of the week (Weekday) and the location in terms of the central area (Centre) seem to have an influence on the frequency of withdrawals. There are noticeable clusters based on these factors, suggesting different spending patterns depending on the day of the week and the location in the city.

3. Models and Methods

The process of developing the final statistical model involved many iterations using a multitude of transformations and methods. Initially OLS was conducted to produce a baseline from which the additional models can be compared with. In addition to OLS, LASSO, Ridge and Elastic Net analysis was performed in attempt to address the risk of multicollinearity highlighted by high correlation between many of the explanatory variables. Each models' predictive capabilities were then accessed in order to ascertain which model was the most successful. Below each method of analysis has been outlined including the preliminary model developed followed by a detailed summary of the selection process and the final model.

3.1. OLS

3.1.1. Overview

Ordinary Least Squares (OLS) is a classical linear regression method used to estimate the parameters in a linear regression model. It aims to minimize the sum of the squared differences between the observed and predicted values. OLS is widely utilized for its simplicity and interpretability and is suitable when the relationship between the independent and dependent variables is approximately linear.

3.1.2. Method

The OLS model is represented as,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Where, Y is the predicted or estimated value of the dependent variable, $X_1, X_2, \dots, X_p$ are the independent variables, β_0 is the intercept, representing the value of Y when all independent variables are zero and $\beta_1, \beta_2, \dots, \beta_p$ are the coefficients that represent the change in Y for a one-unit change in the corresponding independent variable, holding other variables constant.

The OLS method minimizes the residual sum of squares (RSS) to obtain the coefficient estimates. The formula for RSS is:

$$RSS = \frac{1}{n}\sum_{i=1}^{n}\left(Y_{i,observed} - Y_{i,predicted}\right)^2$$

The coefficient estimates $\beta_1, \beta_2, \dots, \beta_p$ are determined using the formula,

$$\beta_{ls} = (X'X)^{-1}X'y$$

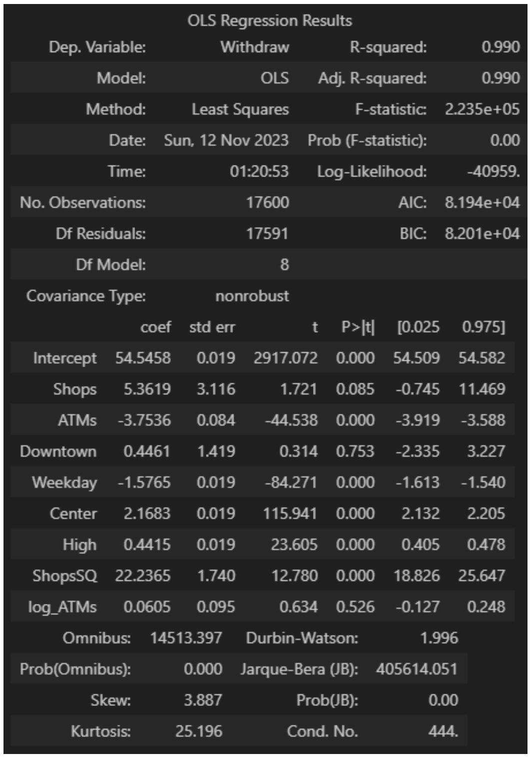
Where, β_{ls} is the vector of estimated coefficients, X is the matrix of independent variables and Y is the vector of observed values of the dependent variable.

This works when the matrix $X'X$ is invertible (i.e., its determinant is non-zero). However, issues can arise when $X'X$ is not invertible or is "near-singular," meaning it has a determinant close to zero. In such cases, problems can be encountered with OLS regression.

The Exploratory Data Analysis highlighted that there is a significant risk of multicollinearity with high correlations between "Shops" and "Downtown", "Shops" and "ATMs"; and "ATMs" and "Downtown”. Thus, a multitude of recession techniques have been used to account for this risk. Furthermore, by accessing the correlations between the output variable “Amount” and the independent variables it has been determined that the introduction of additional transformations could better fit the data and improve model performance. Thus, “Shops” squared and the natural log of “ATMs” (plus a constant to address zero values) have been added to the preliminary model.

3.1.3. Preliminary Model

Performing OLS with the selection of variables results in the following model,



```
                          OLS Regression Results
     Dep. Variable:          Withdraw        R-squared:      0.990
             Model:               OLS   Adj. R-squared:      0.990
            Method:     Least Squares      F-statistic:  2.235e+05
              Date:  Sun, 12 Nov 2023 Prob (F-statistic):     0.00
              Time:          01:20:53   Log-Likelihood:     -40959.
 No. Observations:             17600              AIC:  8.194e+04
      Df Residuals:            17591              BIC:  8.201e+04
          Df Model:                8
  Covariance Type:         nonrobust
                coef   std err          t    P>|t|   [0.025    0.975]
  Intercept  54.5458     0.019   2917.072    0.000   54.509    54.582
      Shops   5.3619     3.116      1.721    0.085   -0.745    11.469
       ATMs  -3.7536     0.084    -44.538    0.000   -3.919    -3.588
   Downtown   0.4461     1.419      0.314    0.753   -2.335     3.227
    Weekday  -1.5765     0.019    -84.271    0.000   -1.613    -1.540
     Center   2.1683     0.019    115.941    0.000    2.132     2.205
       High   0.4415     0.019     23.605    0.000    0.405     0.478
    ShopsSQ  22.2365     1.740     12.780    0.000   18.826    25.647
   log_ATMs   0.0605     0.095      0.634    0.526   -0.127     0.248
       Omnibus:  14513.397    Durbin-Watson:       1.996
 Prob(Omnibus):      0.000  Jarque-Bera (JB):  405614.051
          Skew:      3.887         Prob(JB):        0.00
      Kurtosis:     25.196         Cond. No.        444.
```

While the overall OLS regression models appear to have high explanatory power (R-squared = 0.99) it is immediately obvious that some of the independent variables are not statistically significant. A manual approach to variable selection could be taken, removing the least significant variables one by one and re-running the model. However, as the regressions techniques below act as a method to address the issue on non-significant variables in an automated manner this manual approach will not be used.

3.2. Lasso Regression

3.2.1. Overview

LASSO (Least Absolute Shrinkage and Selection Operator) is a type of linear regression that implements shrinkage. That is the method attempts to shrink the data values towards a central point such as the mean. Thus, this procedure imposes a penalty of complexity and encourages simple, sparse models and hence is well suited for models with a high degree of multicollinearity.

The LASSO procedure uses L1 regularization implementing a penalty on the magnitude of the coefficients. In some cases, large penalties can result in coefficient shrinking to zero and are eliminated from the model. Thus, LASSO can also be considered as an automated selection technique.

3.2.2. Method

Lasso adds a penalty equal to the absolute value of the magnitude of coefficients. Its goal is to minimise the following,

$$\sum_{i=1}^{m}\left(y_i-\sum_{j}^{n} z_{ij}\beta_j\right)^2+\lambda\sum_{j}^{n}|\beta_j|$$

Where $\boldsymbol{y}$ is the vector of response variables, $\boldsymbol{Z}$ is the matrix containing all predictors, $\boldsymbol{\beta}$ the vector of all population parameters to be estimated and λ is the shrinkage parameter that modulates the penalty imposed on the magnitude of coefficients. An increase λ correlates with a greater penalty and bias whilst a decrease correlates with a lesser penalty and greater variance. When λ is equal to zero the LASSO method behaves like a regular OLS.

Before applying this method, the predictors must first be standardised.

$$\sum_{j}^{n} z_j = 0, \qquad \frac{1}{n}\sum_{j}^{n} z_j^2 = 1$$

This can be done by performing the following operation.

$$\boldsymbol{z_i}=\frac{\boldsymbol{x}_i-\overline{\boldsymbol{x}}}{\boldsymbol{\sigma}}$$

Where $\boldsymbol{x}_i$ is the unstandardised ith observation, $\overline{\boldsymbol{x}}$ is the mean of the unstandardised parameters and $\boldsymbol{\sigma}$ is the standard deviation of the unstandardised parameters.

Once this is done a shrinkage parameter can be selected and the algorithm can be performed. Selection of the shrinkage parameter for the preliminary model below was done by using cross validation. Cross-validation is a resampling method that uses different portions of the data to test and train a model on different iterations. In this analysis python's "LassoCV" function was used with 5-fold leave-one-out cross validation to derive the estimators.

3.2.3. Preliminary Model

Using pythons "LassonCV" method with 5-fold leave-one-out cross validation the following shrinkage parameter was found to minimise the error of the model,

$$\lambda = \ 0.024788566404595102$$

This value for the shrinkage parameter results in the following model,

$$\begin{aligned}Withdraw = \beta_0+\beta_1*Shops+\ \beta_2*Shops^2+\beta_3*ATMs+\beta_4*\log(ATMs+1)+\beta_5*Downtown+\beta_6\\ *Weekday+\beta_7*Centre+\beta_8*High\end{aligned}$$

With,

$$\begin{aligned}&\beta_0 = 54.5458 && \beta_5 = 0\\ &\beta_1 = 11.8395 && \beta_6 = -1.5496\\ &\beta_2 = 16.0487 && \beta_7 = 2.1460\\ &\beta_3 = -3.5380 && \beta_8 = 0.4198\\ &\beta_4 = 0 &&\end{aligned}$$

3.3. Ridge Regression

3.3.1. Overview

Ridge regression, like LASSO, is a statistical tuning method that can be used to analyse data suffering from multicollinearity. I utilizes L2 regularisation to impose a penalty upon the magnitude of the model's coefficients. Unlike LASSO the coefficients will only be scaled by the shrinkage parameter not reduced zero and eliminated from the model.

3.3.2. Method

Ridge adds a penalty equal to the square of the coefficients. Its goal is to minimize the following,

$$\sum_{i=1}^{m}\left(y_i - \sum_{j}^{n} z_{ij}\beta_j\right)^2 + \lambda \sum_{j}^{n} \beta_j^2$$

Where $\boldsymbol{y}$ is the vector of response variables, $\boldsymbol{Z}$ is the matrix containing all predictors, $\boldsymbol{\beta}$ the vector of all population parameters to be estimated and λ is the shrinkage parameter that modulates the penalty imposed on the magnitude of coefficients. Like Lasso an increase λ correlates with a greater penalty and bias whilst a decrease correlates with a lesser penalty and greater variance. When λ is equal to zero the Ridge method behaves like a regular OLS.

Like LASSO, the predictors must be standardised before applying the method. Refer to section 3.2.2 for an explicit explanation on how this standardisation is performed.

Once this is done a shrinkage parameter can be selected and the algorithm can be performed. Selection of the shrinkage parameter for the preliminary model below was done by using cross validation. Cross-validation is a resampling method that uses different portions of the data to test and train a model on different iterations. In this analysis python's "RidgeCV" function was used with 5-fold leave-one-out cross validation to derive the estimators.

3.3.3. Preliminary Model

Using pythons "RidgeCV" method with 5-fold leave-one-out cross validation the following shrinkage parameter was found to minimise the error of the model, note that whilst "LassoCV" requires no other inputs the "RidgeCV" method requires a vector of shrinkage parameters to try (a linespace vector was used spanning e^{-20} to e),

$$\lambda = 0.014721638308430757$$

This value for the shrinkage parameter results in the following model,

$$Withdraw = \beta_0 + \beta_1 * Shops + \beta_2 * Shops^2 + \beta_3 * ATMs + \beta_4 * \log(ATMs + 1) + +\beta_5 * Downtown + \beta_6 * Weekday + \beta_7 * Centre + \beta_8 * High$$

With,

$$\beta_0 = 54.5458$$
$$\beta_1 = 5.5218$$
$$\beta_2 = 22.1448$$
$$\beta_3 = -3.7535$$
$$\beta_4 = 0.0604$$

$$\beta_5 = 0.3777$$
$$\beta_6 = -1.5765$$
$$\beta_7 = 2.1683$$
$$\beta_8 = 0.4415$$

3.4. Elastic Net

3.4.1. Overview

Given our dataset's multi-dimensional nature, a refined approach beyond the basic OLS is warranted to account for potential multicollinearity and to enhance the model's predictive performance. To this end, we utilize Elastic Net regression—a hybrid penalized regression method that combines the L1 and L2 regularization of Lasso and Ridge regression techniques, respectively.

Elastic Net is a regularization and variable selection method that linearly combines the L1 and L2 penalties of the Lasso and Ridge methods in regression analyses. Developed to blend the strengths of both the Lasso and Ridge, this method is particularly useful when dealing with highly correlated predictors or when the number of predictors is greater than the number of observations.

3.4.2. Method

Elastic net combines both LASSO and Ridge methods. Its goal is to minimize the following,

$$\sum_{i=1}^{m}\left(y_i-\sum_{j}^{n} z_{ij}\beta_j\right)^2+\lambda\sum_{j}^{n}\left(\alpha\beta_j^2+(1-\alpha)\left|\beta_j\right|\right)$$

Where $\boldsymbol{y}$ is the vector of response variables, $\boldsymbol{Z}$ is the matrix containing all predictors, $\boldsymbol{\beta}$ the vector of all population parameters to be estimated and λ is the shrinkage parameter that modulates the penalty imposed on the magnitude of coefficients. An increase λ correlates with a greater penalty and bias whilst a decrease correlates with a lesser penalty and greater variance. α is the mixing parameter between Ridge (L2) and LASSO (L1): when $\alpha = 0$, the penalty is purely Lasso, and when $\alpha = 1$, it is purely Ridge.

When applying Elastic Net, the process starts with standardizing the predictor variables to ensure they're on the same scale. The coefficients are then iteratively determined using the path wise coordinate descent method. The optimal regularization parameters λ and α are identified through cross-validation, aimed at minimizing the prediction error. Thanks to its L1 penalty, Elastic Net can produce sparse models with few non-zero coefficients, which is particularly helpful for variable selection in the presence of many predictors. A significant advantage of Elastic Net is its ability to handle highly correlated variables and provide a unique solution even when Lasso might offer multiple solutions, making it exceptionally suitable for analysing high-dimensional datasets where the number of features exceeds the number of observations.

In this analysis python’s “ElasticNetCV” function was used with 5-fold leave-one-out cross validation to derive the estimators.

3.4.3. Preliminary Model

Using pythons “ElasticNetCV” method with 5-fold leave-one-out cross validation the following shrinkage parameter was found to minimise the error of the model, note that the method requires test alphas (the mixing parameter to be used), and range from 0.01 to 0.99 was parsed into the function,

$$\lambda = 0.025038955964237462$$

$$\alpha = 0.99$$

This value for the shrinkage parameter results in the following model,

$$Withdraw = \beta_0 + \beta_1 * Shops + \beta_2 * Shops^2 + \beta_3 * ATMs + \beta_4 * \log{(ATMs + 1)} + +\beta_5 * Downtown + \beta_6 * Weekday + \beta_7 * Centre + \beta_8 * High$$

With,

$$\beta_0 = 54.5458 \qquad \beta_5 = 0$$
$$\beta_1 = 11.4209 \qquad \beta_6 = -1.5493$$
$$\beta_2 = 16.4485 \qquad \beta_7 = 2.1454$$
$$\beta_3 = -3.5215 \qquad \beta_8 = 0.4196$$
$$\beta_4 = 0$$

3.5. Selection Process

During the derivation of the preliminary model the data was split into a training and validation set, many sources recommend a 80:20 split of data thus this implemented. The training set was used in conjunction with the above methods to estimate the model. These methods use the training error in conjunction with other factors to estimate the population coefficients.

To select the optimum model from the preliminary models above the prediction error was used. The prediction error is derived from the validation set and is calculated as follows,

$$Predictor\ Error = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \sum_{j}^{n} x_{ij}\beta_j\right)^2}$$

Where $\boldsymbol{y}$ is the vector of response variables of the test set, $\boldsymbol{X}$ is the matrix containing all predictors for the test set, $\boldsymbol{\beta}$ the vector of all population parameters estimated by the chosen method and m is the number of observations in the test set. Simply it is the average of the difference between the actual value and the predicted value.

Taking the average prediction error for each model it is possible to then compare models' accuracy. As such the model with the lowest prediction error, that is the model that can explain the validation sets data with the most accuracy, is chosen as the final model.

The predictor error for the 4 models explored above are as follows,

Method	Predictor Error
OLS	2.522586
Lasso	2.532416
Ridge	2.522672
Elastic Net	2.532015

Note that although OLS performed best resulting in the lowest predictor error this model will not be chosen as the final model. The explanatory data analysis in section 2 outlined the high risk of multicollinearity and hence various regularisation techniques have were implemented. Whilst the shrinkage factors of each method are close to zero they importantly have not converged towards this point. Namely the ridge regularisation method given the option to select e^{-20} (2.06115362e-9) as the shrinkage parameter, determined through cross validation that a greater shrinkage parameter and hence penalty on the model's complexity performed better. This indicates that multicollinearity is a necessary factor to be considered and thus OLS is an inappropriate choice. Had the data been re-randomised and the analysis re-run it is likely that the different selection of data would result in OLS producing a higher predictor error.

Thus, the ridge model has been chosen as the final model as smallest prediction error excluding OLS.

3.6. Final Model

$$Withdraw = \beta_0 + \beta_1 * Shops + \ \beta_2 * Shops^2 + \beta_3 * ATMs + \beta_4 * \log{(ATMs+1)} + \ +\beta_5 * Downtown + \beta_6 * Weekday + \beta_7 * Centre + \beta_8 * High$$

With,

$$\beta_0 = 54.5458 \qquad \beta_5 = 0.3777$$
$$\beta_1 = 5.5218 \qquad \beta_6 = -1.5765$$
$$\beta_2 = 22.1448 \qquad \beta_7 = 2.1683$$
$$\beta_3 = -3.7535 \qquad \beta_8 = 0.4415$$
$$\beta_4 = 0.0604$$

4. Discussion

Comparing the Lasso model with the Ridge model, we note that the Ridge regression has a slightly lower prediction error than the Lasso (2.522672 for Ridge vs. 2.532416 for Lasso), according to the provided data. While both models add regularization terms to minimize overfitting, Ridge regression tends to keep all variables in the model but shrinks the coefficients towards zero to reduce model complexity. This may suggest that in this particular case, the Ridge model is better suited when we want to retain all features and deal with potential multicollinearity, assuming the variables are relevant. It also indicates that the Lasso's feature elimination approach might be too aggressive, potentially disregarding valuable information that the Ridge model retains, which could be why Ridge has a lower prediction error and might be preferred over Lasso.

The Elastic Net regression model used here, tuned with a 5-fold leave-one-out cross-validation approach, combines L1 and L2 regularization to enhance predictive accuracy while managing multicollinearity and feature selection. The optimal parameters indicate a strong preference for feature selection, given the high alpha value of 0.99. The model's selected coefficients suggest that certain predictors such as the number of shops have a nonlinear effect on the withdrawal amounts, while others like ATMs' presence are negatively correlated or excluded from the model, like the "Downtown" variable. However, in model comparisons, its prediction error is not the lowest. Prediction error is a key metric for assessing the predictive ability of a model. If other models, such as ordinary least squares, Lasso, or Ridge regression, provide lower errors, it suggests that they may be better suited for the dataset and can predict the outcomes more accurately.

the Ordinary Least Squares (OLS) model, despite yielding the lowest prediction error, is not selected as the optimal model. This is due to OLS's inability to manage multicollinearity, which can lead to inflated variance in the coefficient estimates and reduce the model's generalizability. Regularization methods like Ridge regression, on the other hand, introduce a penalty term that constrains the coefficient estimates, thus providing a solution to the multicollinearity problem. The choice of a non-trivial shrinkage factor for Ridge, as determined through cross-validation, reflects its effectiveness in handling the complexities of the data. This regularization process results in a model that, while potentially having a slightly higher prediction error than OLS on the current dataset, is expected to exhibit more stable performance on different datasets. This stability is crucial, as OLS's performance is likely to degrade with a new randomization of the data, indicating that a regularized model like Ridge is a better choice for this scenario.

In the case of linear models like Ridge Regression, the coefficients can provide insights into the impact of each feature on the target variable. The coefficients represent the change in the dependent variable for a one-unit change in the corresponding independent variable, holding other variables constant. The introduction of the penalty term can shrink the coefficients, making interpretation slightly more complex. However, it still allows for an understanding of the direction and approximate magnitude of the relationships between the features and the target variable.

In conclusion, while Ridge Regression helps address multicollinearity and improve model performance, it is important to be aware of potential pitfalls such as overfitting, data assumptions, and the complexity introduced by the regularization process. Balancing model interpretability and performance is crucial, and careful consideration of these factors is necessary when interpreting the results and making informed decisions based on the analysis. Regular validation and sensitivity analyses are essential to ensure the robustness and reliability of the chosen model's performance.

5. References

- GeeksforGeeks. (2020). Elastic Net Regression in R Programming. [online] Available at: https://www.geeksforgeeks.org/elastic-net-regression-in-r-programming/.
- Brownlee, J. (2020). How to Develop Elastic Net Regression Models in Python. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/elastic-net-regression-in-python/.
- www.sthda.com. (n.d.). Penalized Regression Essentials: Ridge, Lasso & Elastic Net - Articles - STHDA. [online] Available at: http://www.sthda.com/eng
- Oup.com. (2023). Available at: https://academic.oup.com/jrsssb/article/67/2/301/7109482?login=true.
- Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society. Series B (Statistical Methodology), [online] 67(2), pp.301–320. Available at: https://www.jstor.org/stable/3647580.
- Godwin, J.A. (2021). Ridge, LASSO, and ElasticNet Regression. [online] Medium. Available at: https://towardsdatascience.com/ridge-lasso-and-elasticnet-regression-b1f9c00ea3a3.
- Jain, S. (2023, May 1). *Lasso & Ridge Regression | A Comprehensive Guide in Python & R (Updated 2023)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/
- Stephanie. (2021b, April 27). *Lasso Regression: Simple Definition - Statistics How to*. Statistics How To. https://www.statisticshowto.com/lasso-regression/
- Stephanie. (2023b, March 9). *Ridge Regression: Simple Definition - Statistics How To*. Statistics How To. https://www.statisticshowto.com/ridge-regression/
- Great Learning Team. (2022b, November 16). *Ridge Regression Definition & Examples | What is Ridge Regression?* Great Learning Blog: Free Resources What Matters to Shape Your Career! https://www.mygreatlearning.com/blog/what-is-ridge-regression/#:~:text=Ridge%20regression%20is%20a%20model,away%20from%20the%20actual%20values.
- Great Learning Team. (2023, May 30). *What is LASSO Regression Definition, Examples and Techniques*. Great Learning Blog: Free Resources What Matters to Shape Your Career! https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/#:~:text=Lasso%20regression%20is%20a%20regularization,i.e.%20models%20with%20fewer%20parameters)
- *sklearn.linear_model.LassoCV*. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LassoCV.html
- *sklearn.linear_model.RidgeCV*. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.RidgeCV.html
- Melkumova, L. E., & Shatskikh, S. Y. (2017). Comparing Ridge and LASSO estimators for data analysis. *Procedia Engineering*, *201*, 746–755. https://doi.org/10.1016/j.proeng.2017.09.615