12 November 2021

UNSW
SYDNEY

Group Assignment

ECON3203: Econometric Theory and Methods

PRESENTED BY

Dharani Palanisamy (z5260276)
Dharshini Loganathan (z5309765)
Faiyam Islam (z5258151)
Mimi Nguyen (z5260968)

Introduction

The objective of this report is to develop a predictive model for ATM cash demand. As employees of the bank, we are required to optimise the bank's cash management system by permeating smarter decisions in reloading its ATM network. In this investigation, we have relied on the data collected from the ATM training dataset. It is imperative that an understanding of predictive accuracy is established in order to utilise the ideal covariates. The dependent variable is the total cash withdrawn a day, *Withdraw* which is explained by various covariates related to the bank. Simply stating that a particular subset of independent variables best explains the outcome will not suffice. Thus, various models with multiple estimation methods and statistical inferences will be considered, in conjunction with model selection, classification and neural network methods to analyse and optimise this bank's cash management.

Cash management is a complicated task due to the interaction between multiple monetary activities such as investments and funding (Moubariki, 2019). In order to avoid major financial consequences, accurate machine learning and deep learning techniques need to be correctly utilised. Predicting cash demand is challenging due to the unpredictability of withdrawals, however developing advanced algorithms can improve returns on cash assets and lower operational costs simultaneously (Visionet, 2021). Forecasting amounts of cash withdrawn daily requires a certain combination of independent variables and in order to propagate predictive analysis, we will be using the test error computed as follows:

$$\text{Test\_error} = \frac{1}{n_{\text{test}}} \sum_{y_i \in \text{test data}} (\widehat{y}_i - y_i)^2$$

where n_{test} is the number of observations in our test data. Within our analysis, we attempt to determine the most suitable model which can remediate the problem of obtaining a high test error.

Initial analysis of our dataset will consist of an exploratory data analysis in understanding the different variables followed by simple and multiple linear regression which will produce credible estimations of these models. This will be followed by regularisation models such as Ridge and LASSO which will allow us to counteract multicollinearity issues. In addition, to directly address the smallest possible prediction error, classification models such as K-nearest neighbours (KNN) classification intertwined with cross validation and neural network techniques will be incredibly useful. Through model selection, the determination of a subset of covariates will allow us to implement the smallest possible test error. However, there are a myriad of limitations to many of our methods which will be discussed towards the end of this report as well as an evaluation of the test errors of the models.

Exploratory Data Analysis (EDA)

An initial exploration of the data reveals a total of 6 covariates, consisting of a combination of continuous and categorical covariates, and variable *Withdraw*, which strongly indicates that discrete models and classification methods will not be appropriate for this data. The *Withdraw* variable will be the dependent variable in our analysis. We also observe that there are a total of 22,000 data points, all of which are non-null. The maximum amount withdrawn is observed to be $103,964 for this sample, and the minimum only $1167. Furthermore, there are more ATMs in downtown areas, for most ATMs, there were no high demands for cash the previous month.

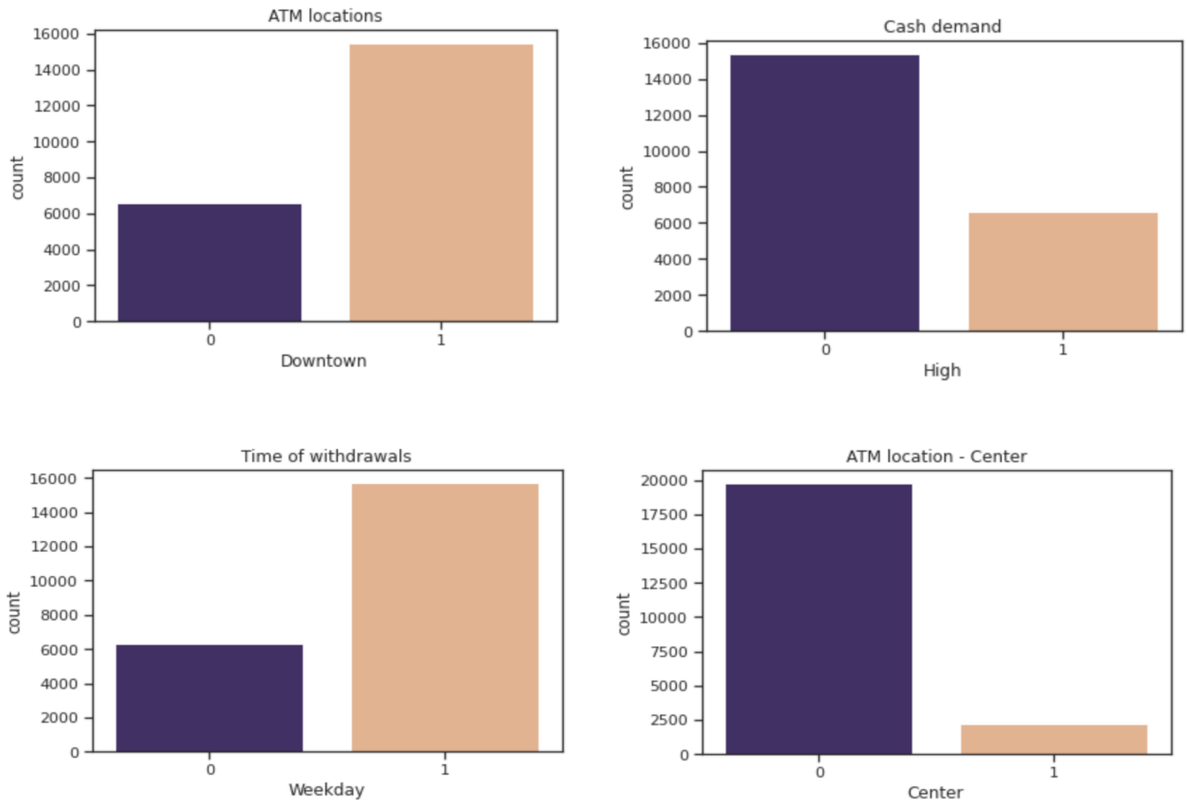
We observe multimodal distributions in all the continuous covariates and the outcome variable, as shown in the histograms in Figures 1, 2 and 3, and further reinforced in the pairplot of Figure 4 . For *Withdraw*, this strongly indicates that there are three commonly withdrawn amounts, specifically around 20,000, 70,000, and 90,000 in the local currency, and it is likely that these are the withdrawal limits corresponding to the varying cash demand in different ATM locations. As such, this strongly reveals that the normal distribution does not hold for the sample data, which suggests that the least squares model will not be a good fit.

Moreover, we observe high collinearity between some of the covariates in the correlation matrix of the heatmap in Figure 5. Specifically, there is strong collinearity between *Shops* and *ATMs*, and *Downtown* and ATMs, and perfect collinearity between *Shops* and *Downtown*. Thus, this reinforces that OLS estimation will not yield optimal predictions. Interestingly, it is observed that the binary covariates – *High, Center,* and *Weekday* – have negligible correlation with the *Withdraw*, and thus it could be hypothesised that there is no location nor time effect on cash demand, and as such, we can expect to remove these covariates from our models. Furthermore, the signs of these correlation coefficients align with our hypotheses of how these covariates relate to *Withdraw*, by which they all positively correlate with *Withdraw*, except for *Weekday*. It is possible that a majority of card holders are more busy during the weekday, and are therefore less likely to *Withdraw* than during the weekend.

An initial exploration of the data reveals a total of 6 covariates, consisting of a combination of continuous and categorical covariates, and dependent variable *Withdraw*, which strongly indicates that classification methods will not be appropriate for this data. Some summary results of the dataset exemplifies no null values and duplicates with a maximum withdrawal of $103,964 and minimum withdrawal of $1167.

Initial analysis of the existing categorical variables reveals that there are more records of ATM’s in the downtown areas and for the majority of ATMs, there were no high demands for cash the previous month. Additionally, more ATM’s have recorded withdrawals over the weekdays compared to weekends. However, it should be noted that there are more weekdays than weekend days, thus inferring that withdrawals are roughly similar on any given weekday and weekend in comparison. Majority of ATM’s in the dataset are not located within a center.

Figure 1: Boxplot of covariates against Withdraw variable



Initial exploration of the existing numerical variables renders that the number of shops (in 100s) within walking distance to an ATM is either very low or very high. This can mean the location is either isolated, or in a commercial location. The number of ATMs (in 10s) within walking distance holds a multi-modal distribution.

Figure 2: Histogram of Shops and ATMs

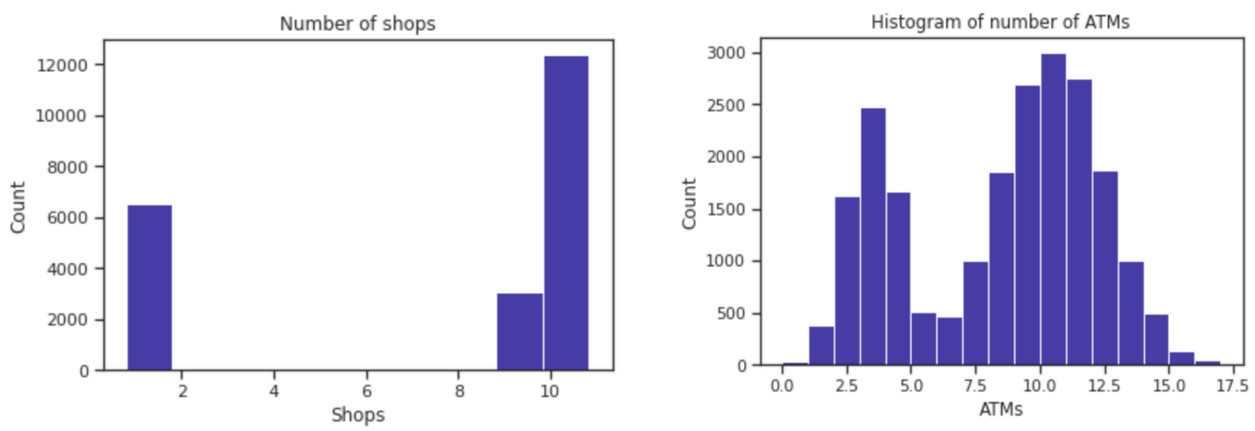


Figure 3: Scatterplot showing relationship between covariates and Withdraw

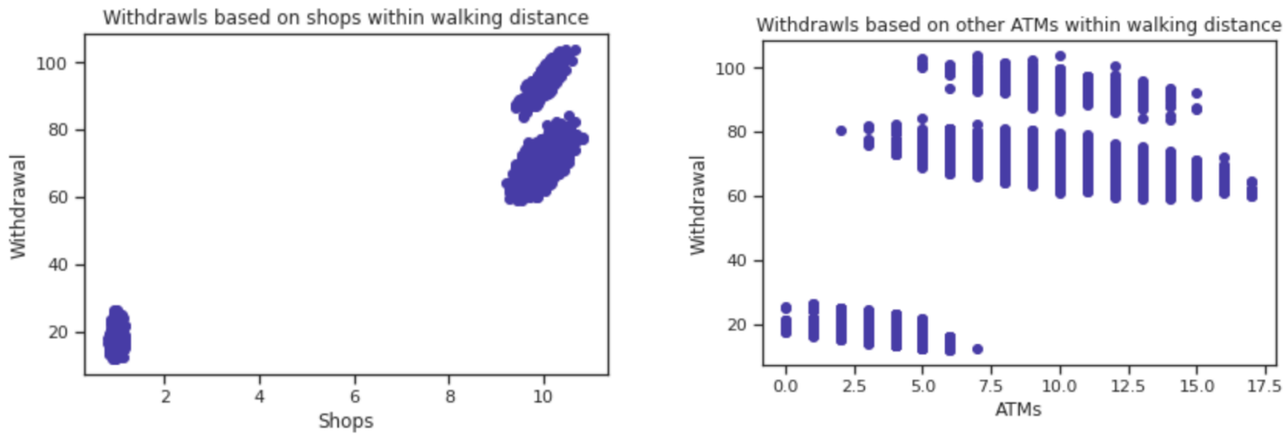
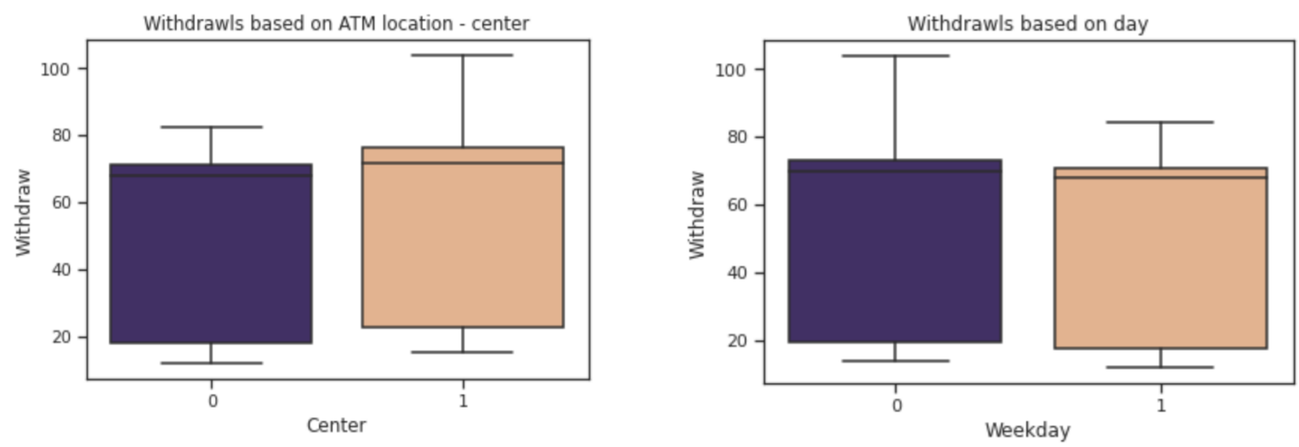


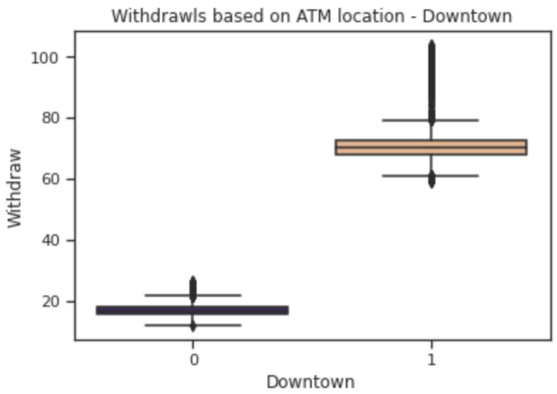
Figure 3 strongly indicates that there are three commonly withdrawn amounts, specifically around 20,000, 70,000, and 90,000 in the local currency, and it is likely that these are the withdrawal limits corresponding to the varying cash demand in different ATM locations. As such, this strongly reveals that the normal distribution does not hold for the sample data, which suggests that the least squares model will not be a good fit.

Relationship between categorical covariates and *Withdraw*

There is a clear distinction between withdrawal amounts for ATM's which are located downtown and those which are not. ATMs which are located downtown see a much larger total cash amount withdrawn daily compared to ATMs which are not located downtown. The independent variable *Center* also seems to have a noticeable effect on withdrawals made from the bank. The maximum and average withdrawals made from an ATM located in a center is greater than withdrawals made from an ATM not located in a center. This relationship could be attributed to the fact that there are potentially more shops around the area within a center. Withdrawals during weekdays and weekends are approximately similar. However, weekends have a noticeably higher maximum withdrawal amount. Furthermore, the interquartile range of withdrawals is mostly identical for ATMs with high or lower cash demand in the previous month. From Figure 4 (High against Withdraw), this covariate seems to have the least effect on withdrawal and it is likely that the huge discrepancies between withdrawal amounts are mostly correlated to other factors such as *Shops*, *Downtown* and *ATMs*.

Figure 4: Boxplots of covariates against Withdraw variable





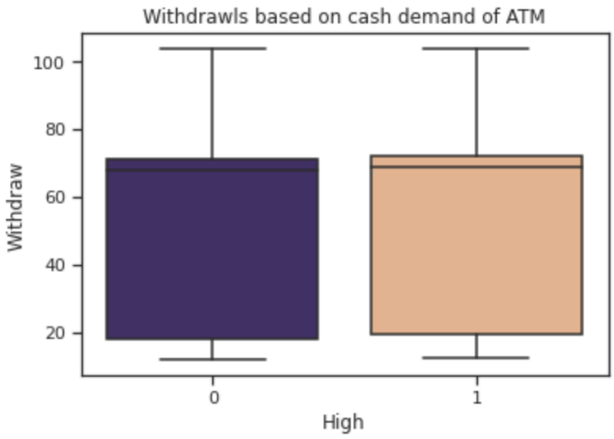


Figure 5: Heatmap of correlation matrix

We observe high collinearity between some of the covariates in the heatmap correlation matrix in Figure 5. Specifically, there is strong collinearity between *Shops* and *ATMs*, and *Downtown* and *ATMs*, and perfect collinearity between Shops and *Downtown*. Thus, this reinforces that OLS estimation will not yield consistent estimates, and thus optimal predictions. Interestingly, it is observed that the binary covariates - *High*, *Center*, and *Weekday* – have negligible correlation with the *Withdraw*, and thus it could be hypothesised that they have a negligible effect on *Withdraw* and we can expect to remove these covariates from our models. Furthermore, the signs of these correlation coefficients align with our hypotheses of how these covariates relate to *Withdraw*, by which they all positively correlate with *Withdraw*, except for Weekday. It is possible that a majority of card holders are more busy during the weekday, and are therefore slightly less likely to *Withdraw* cash on a weekday as opposed to the weekend. As analysed earlier, weekends also have a noticeably higher maximum withdrawal amount. The relationships mentioned here are reinforced in the pairplot shown [Appendix A].

Methodology

A range of analytical techniques have been implemented, including regression analysis, regularisation, classification and deep learning methods. The incentive is to train and test the relationship between the response variable and covariate variables to measure the prediction accuracy in which we shall be using the mean squared error for reference.

1. Simple and Multiple Linear Regression

Simple Linear Regression (SLR) is a useful and appropriate approach to adapt when predicting a quantitative response to a single predictor variable. The residual sum of squares (RSS) is the difference between the ith observed response value and the ith response value. The least squares (LR) method chooses the coefficients which minimizes the RSS and will obtain us with least square coefficient estimates for LR.

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

or equivalently as

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

From initial exploratory analysis, the variables *Shops*, *ATMs* and *Downtown* seemed to produce biggest variations within *Withdraw*. An investigation of the effects of the independent variables on *Withdraw* was carried out by regressing *Withdraw* against each of the covariates. The results of the goodness of fit are as given below:

	Shops	Downtown	ATMs	Center	Weekday	High
R^2	0.972	0.967	0.679	0.008	0.003	0.000

As hypothesised from the EDA, the variables *Shops*, *ATMs* and *Downtown* explain the majority of the variation within *Withdraw*. Unlike the simple linear regression estimates, multiple regression coefficient estimates are commonly represented using matrix algebra and thus have complicated forms. We experiment with multiple predictors to improve our fit and predictions and target underfitting issues that were previously seen with SLR. Cross validation with 10 folds was used again in an attempt to resolve possible overfitting. The cross validation scores obtained from each of the 10 folds will be averaged to generate a final mean squared error (MSE). An analysis of the effects of the independent variables on *Withdraw* was carried out by regressing *Withdraw* with all of the covariates to create a base model.

	Shops	Shops, ATMs	Shops, ATMs, Center	Shops, ATMs, Center, Weekday	Shops, ATMs, Center Weekday, High
Adjusted R^2	0.972	0.978	0.982	0.990	0.990

In essence, the base model becomes:

$$Withdraw = \beta_0 + \beta_1 Shops + \beta_2 ATMs + \beta_3 Downtown + \beta_4 Center + \beta_5 Weekday + \beta_6 High$$

At first glance, MLR seems to be performing much better than the SLR model. Observe how as the number of variables increase, consequently the adjusted R^2 value does too. In relation to the SLR model, when weak predictors are added to

the stronger covariates, more of the variation is explained with the dependent variable, *Withdraw*. Overall, it can be seen that adjusted R^2 values in the table above, are larger in comparison to when the strong predictors are estimated singularly, and multiple linear regression seems to be more relevant.

From the table below, we can see that the singular effect of *Downtown* hugely impacts on *Withdraw* based on previous SLR analysis, than if it were to be regressed in combination with its *Shops* which it shares extremely high collinearity with. Statistical consequences of multicollinearity mean that we may be unable to declare an independent variable significant even though by itself it has a strong relationship with Y. In our case, while *Downtown* has a noticeable impact on *Withdraw* standalone, regressing it along with *Shops* does not make a significant enough impact on the fit of the model. In fact, since most of the variation on *Withdraw* explained by *Downtown* is actually done by *Shops*, the variable *ATMs* has a larger combined effect on the *Withdraw* variable, slightly different to what we initially hypothesised off the SLR model.

	Shops	Shops, Downtown
Adjusted R^2	0.972	0.973

Henceforth, a brief analysis of the effect of the variable *Downtown* on the MSE was explored due to the rise in multicollinearity resulting from its inclusion in an MLR model. Furthermore, the effect of the variable *High* was also explored. As seen in the table below, the variable “High” does not improve on the fit of the model in any way and henceforth can possibly be discarded in our final MLR model.

	Shops	Shops, High
Adjusted R^2	0.972	0.972

Further exploration of the effect of these covariates will render the predictive MSE in retaining a more conclusive result. Interaction terms can be implemented to extend the model by adding another predictor, which is constructed by computing the product of two or three parameters. When one or more variables are dependent on another, multiplying these terms, then through additive modelling will most likely show superiority to a model that only contains main effects. Since interaction variables have a direct correlation with the original variables they arise from, the data has been standardised yet again to reduce structural multicollinearity that may arise with this. Several interaction terms have been and we will try a few of these interactions as explored below, reasoning our choices intuitively as we proceed. The interaction variable ‘Center_Weekday_Shops’ was chosen. At initial glance while *Weekday* and *Center* seemed to have a negligent standalone effect on *Withdraw*, we hypothesised it was possible for it to have a combined effect with *Center* and *Shops*. It is possible that the effect that shops within walking distance of a local ATM have on the amount withdrawn is directly correlated to the location of the ATM. Since “walkable distance” is subjective in definition, it is highly likely that the amount withdrawn can be higher if ATMs in a center (Center = 1) had numerous surrounding shops compared to if the ATM was located outside of a center (Center = 0) with just as many shops but located further from the ATM (although still within walking distance). Furthermore, the effect that Shops has on withdrawal can fluctuate depending on whether the

current day is a weekday or weekend. Shops can be extremely busy during weekends which can result in a higher amount withdrawn during this time of the week compared to if it were a weekday. Behind this possible logic, we have developed a model using the interaction involving these three terms which could possibly have a joint effect on the amount withdrawn, exploring the marginal effect of *Shops* on *Withdraw* moderated by the location of the ATM and the time of the week. The relevant two way interaction terms are also included within the regression.

Simple and multiple linear regression has been included in our analysis to understand if all the predictor variables help to explain the dependent variable. However, it is questionable whether linear regression produces expected results for out of sample predictions due to some unsatisfied conditions of using SLR and MLR. A discussion of parameters that need possible exclusion and the requirements that need to be met in order to be able to use SLR will be discussed in further detail below. We will use these to assess the accuracy of our prediction in comparison to other regression techniques that are likely able to perform with higher scope.

Ridge Regression and Elastic Net

Incongruous to the least square method, the coefficients are estimated by minimising with a tuning parameter $\lambda \geq 0$.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2$$

Ridge regression is more desirable compared to least squares because of the bias-variance trade off. As the shrinkage parameter λ increases, the flexibility of the ridge regression model is decreased thus leading towards more bias and less variance. To address and counteract multicollinearity issues, Ridge regression has been implemented in our analysis as a regularisation method.

Elastic net, which is a compromise between ridge and LASSO will perform variable selection and shrink the coefficients of correlated predictors like ridge regression. Unlike ridge regression, elastic net utilises the function:

$$\widehat{\beta}_{\text{EN}}=\underset{\beta}{\text{argmin}}\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left(\alpha\beta_j^2+(1-\alpha)|\beta_j|\right),$$

for $\lambda \geq 0$ and $0<\alpha<1$.

2. LASSO Variable Selection

A major disadvantage of the Ridge regression model is that unlike variable selection methods like forward and backward stepwise selection, Ridge regression involves all predictors in the final model. This is an issue due to the added challenges in prediction accuracy, however LASSO solves this disadvantage. LASSO coefficients are determined by minimising the quantity:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|$$

LASSO will be an invaluable analytical technique in eliminating certain covariates as the penalty term forces certain covariates to zero when λ is large.

3. KNN Regression

Employing KNN regression, we can average the observations in the data to approximate the associations and relationships between the covariates and *Withdraw.* The driving mechanism of this method is determining the optimal value for K nearest neighbours for the average of the label. To achieve this optimal K value which will yield the lowest MSE, we implement cross-validation to determine the K value from a range between 1 and 50. Moreover, we can expect our KNN model to perform optimally given the low-dimensionality of our data. The K-nearest neighbours classifier will be used to estimate the conditional probability for the training sample from the training dataset. The conditional probability will be calculated with the most optimal value of K which will be determined through cross validation. Different combinations of predictors were constructed according to domain knowledge, the exploratory data analysis, and results from other models. Initially, only the continuous variables were selected for fitting the model in accordance to their high correlation with *Withdraw* from Figure 5 showcasing the heatmap of the correlation matrix of covariates. Further individual inclusions of the categorical predictors revealed *Downtown* to be the only covariate with minimal effect on lowering the prediction error.

4. Neural Networks

Neural networks are a multilayered assembly of units which send information over weighted connections. Similar to neurons, a neural network consists of an input, hidden and output layer. Through the construction of these networks will provide us with a functional approximation and subsequently allow us to predict the mean squared error of a set of covariates. Forward propagation algorithm will be utilised in determining an output based on a subset of the input layer and hidden layers. It is imperative to note that the number of layers and neurons as well as the type of optimiser will need to be tuned in order to achieve the lowest possible test error.

Results

Simple and Multiple Linear Regression

Computing mean squared errors from the base SLR model with the specified covariates below in Table 1

Table 1: Mean Squared Error of various interaction terms in MLR model

	Shops	Downtown	ATMs
MSE	17.7686	20.5269	202.2269

As explored in the methodology, this model is likely under-fitted since there are a lot more explanatory variables that likely have a significant effect on the dependent variable *Withdraw*.

Running some MLR models discussed in the methodology, we get MSE results as below in Table 2

Table 2: Mean Squared Error of various interaction terms in MLR model

	Shops + ATMs + Downtown + Center + Weekday + High	Shops + ATMs + Center + Weekday + High	Shops + ATMs + Center + Weekday
MSE	6.2578	6.7306	6.9230

The MLR has a much higher predictive performance in comparison to the SLR model as hypothesised. The base MLR model containing all explanatory variables seems to be performing best using cross validation technique. We will now compare this base model against several interaction based linear models to check for performance optimisation. Two way and three way interaction terms have been implemented.

Table 3: Mean Squared Error of various interaction terms in MLR model

	Shops + ATMs +Weekday, Downtown + High +Center, Center_Weekday + Shops_Center + Shops_Weekday + Center_Weekday_Shops	Shops + ATMs + Weekday, Center + Center_Weekday + Shops_Center + Shops_Weekday + Center_Weekday_Shops	Shops + ATMs + Downtown + Weekday + High + Center + Center_Weekday
MSE	0.3108	0.9994	2.366

With the addition of several interaction terms using two to three predictors which are dependent on each other we obtained more desirable results. As shown in table 3, to explain the total *Withdrawals* in a day, the number of *Shops* which are within a walkable distance to a customer is related to whether an ATM is located in a *Center* and this ultimately depends on if the customer withdraws on a weekend or *Weekday*. Hence, when modelling our interaction terms, we made sure to include more relevant predictors which will give us the most interpretable and causing results. After grouping and including such terms to the MLR, there is a further large decrease in the MSE value from our base models. Our additive regression vastly improved giving a MSE of 0.3108.

While additional terms within the dataset decreased MSE further, this comes at the cost of including perfectly collinear variables as well as some other highly multicollinear variables. This will be further addressed in our discussion later.

Ridge Regression and Elastic Net

Table 4: Estimated coefficients of selected models at optimal values of lambda

	Optimal parameter	Shops	ATMs	Downtown	Weekday	Center	High
LASSO	0.0247	27.9023	-3.6249	-0.0	-1.5515	2.1591	0.4157
Ridge	4.5400e-05	44.5388	-3.7088	-16.5527	-1.5820	2.1813	0.4390

Elastic Net	0.0250	27.8697	-3.5956	0.0	-1.5510	2.1586	0.4158

The first regularisation method that has been implemented is Ridge regression. Prior to our analysis of Ridge regression, the covariates have been standardised to maintain consistency. As displayed in Table 4, the optimal Ridge parameter λ is 4.5400. From the estimated coefficients, we discern that the variables *ATMs*, *Downtown* and *Weekday* are negative, whereas *Shops*, *Center* and *High* are positive. Based on the magnitudes of these estimated coefficients, it is suggested that the predictors *Weekday*, *Center*, and *High* do not have a significant economic impact on *Withdraw*. Similarly, elastic net is implemented with a grid of values on ridge penalties. Here, *ATMs* and *Weekday* variables are negative and *Shops*, *Center* and *High* are positive with *Downtown* being 0.0, indicating a removal of this predictor. The resulting L1_ratio attribute of 0.9900 strongly suggests that the penalty favours the L1 penalty, and thus the LASSO implementation.

LASSO Variable Selection

Results from the LASSO output differ greatly to that of the Ridge regression as shown in Table 4. The LASSO coefficients with the shrinkage parameter $\lambda = 1$ conveys that the variables ATMs, Downtown and High are zero, suggesting that these variables should be removed to minimise the prediction error [Appendix C]. Using $\lambda = 10$, we interpret all the predictors to be 0 except Shops [Appendix C]. Cross Validation allows us to select the optimal λ value which minimises the predictive error. However, the estimated coefficients are very similar to that of the elastic net estimates, which is expected given that the L1_ratio attribute yielded is 0.9900. Table 4 displays the optimal LASSO lambda to be 0.0247, resulting in only *Downtown* being the predictor to be of value 0 which is the same as the elastic net method. With a CV MSE of 6.7418, the LASSO is more accurate than elastic net but has a greater error compared to ridge regression, thus concluding that ridge renders the lowest test error.

K-Nearest Neighbours (KNN) Regression

Table : Cross Validation MSE of KNN models with different covariate combinations

	Shops	ATMs	Downtown	Weekday	Center	High	CV MSE
K = 49	✔	✔	-	-	-	-	3.7379
K = 49	✔	✔	✔	-	-	-	3.7316
K = 41	✔	✔	-	✔	-	-	3.3625
K = 44	✔	✔	-	-	✔	-	2.9433
K = 46	✔	✔	-	-	-	✔	3.7245
K = 9	✔	✔	-	✔	✔	-	0.7521
K = 50	✔	✔	-	-	✔	✔	2.9224
K = 5	✔	✔	-	✔	✔	✔	0.6101
K = 50	✔	✔	✔	-	✔	✔	2.9018
K = 16	✔	✔	✔	✔	✔	-	0.7153

K = 12	✔	✔	✔	✔	✔	✔	0.7153

Ticks represent that the covariate has been used

Table 5 above shows the different specifications for the KNN regression model and their corresponding cross-validation MSE (CV MSE). Whilst *Shops* and *ATMs* have been shown to have high correlations with *Withdraw* from the EDA, they are not sufficient on their own to provide accurate predictions of the cash amounts withdrawn. We observe similar MSE's for models which include other variables individually, with the exception of including *Weekday* and *Center.* This specification is where we observe significant decreases in their respective CV MSE's, signifying the potential great impact of these covariates on improving predictions. This is further reinforced when we fit the model using a subset of *Shops, ATMs, Weekday,* and *High*, and obtain a CV MSE of 0.7521 which is a significant drop from our previous lowest error value. However, the lowest CV MSE of 0.6101 corresponds to a K value of 5 and a combination of predictors which are *Shops, ATMs, Weekend, Center,* and *High.* This aligns with the results from LASSO and Elastic Net regressions, and further reinforces the negligible effect of *Downtown* on *Withdraw*.

Neural Networks

Neural networks, which are a set of flexible non-linear methods for regression, classification and other tasks have been executed to obtain our test error. Prior to obtaining the Feed Forward Neural Network output, the data needs to be scaled in preparation of using an activation function which in this case rectified linear unit, or 'relu' was used. Our Feed Forward network uses 1 layer of 12 neurons and using the Adam optimiser, the mean squared error we achieve was 0.2550. The overall incentive of using neural networks was to take advantage of the extremely flexible structure of this machine learning technique. Ultimately, in comparison to other models utilised in our analysis, the test error is relatively smaller. This may be a result of overfitting the model which neural networks are responsible for, causing a high level of bias. In consequence, the complexity and unexplained behaviour of neural networks elevates the difficulty of obtaining a lower mean squared error, however through some tuning we achieved a MSE of 0.2550.

Table 6: Cross Validation Mean Squared Error of all analytical methods utilised

	Simple Linear Regression	Multiple Linear Regression	Ridge Regression	LASSO	Elastic Net	KNN Classification	Neural Networks
CV MSE	17.7686	0.3108	6.2583	6.7418	6.7451	0.6101	0.2550

Table 6 above showcases the test errors of the analytical methods aforementioned. From here we deduce that the smallest test error achieved is from the neural networks model of 0.2550. This will be our selected model. The neural networks model has the lowest CV MSE which is followed by the multiple linear regression model with interactions and KNN regression. This will be our selected model which will be implemented using the test data.

Discussion

The SLR model was inherently not ideal as the model was too simple and had very few parameters. In considering the bias-variance decomposition, the model likely was high bias and low variance. While the MLR involving interaction terms generated reasonably low MSE values, we will ensue a discussion on the possible issues and difficulties of using this model. In conducting EDA, we can see that the effect of shops and other ATMs on withdrawal amount at a given local ATM does not exactly follow a linear relationship. Withdrawal amounts seem to occur more prominently as clusters of values. Thus, the linearity condition is not met and could possibly have given rise to biased results and coefficients. In addition to this, the distribution of withdrawals does not follow a normal distribution, indicating that the distribution of residuals are not normal either. This is yet another unsatisfactory condition when beginning to consider linear regression models for analysis and can have an impact on predictability of new data.

Briefly touching upon the issue of multicollinearity, we can draw from previous exploratory analysis that *Shops* had a perfect collinear relationship with *Downtown*. However, lower values of MSE's were generated when including the collinear variables. Given an adjusted R^2 value of 1, there is a possibility that the high number of explanatory variables included in the final MLR model being fit is trying to account for the random errors in the dataset, and therefore overfits it. The model is likely in this case to be low in bias but high amounts of variance. Therefore, although the test data possesses similar characteristics to the training set and is a random subset of the entire dataset, we run the risk that the model might be unstable on the test data and might produce imprecise results on the predictions run against the test dataset.

LASSO variable selection is an ideal regularisation method that is utilised to counteract issues with overfitting the data. By adding a penalty term, the best fit is derived from the trained dataset. The MLR models comprising of certain interaction terms was disadvantaged by the overfitting issue with too many covariates in the model resulting in inflated R^2values. The main issue contrived from our analysis of LASSO is the omission of certain variables that are significant in our analysis. Aforementioned in the results, when $\lambda = 1$, the covariates ATMs, Downtown and High are removed [Appendix C]. Although this allows LASSO coefficients to be more interpretable, the predictive power is decreased due to lack of covariates. Similarly can be explained when $\lambda = 10$ [Appendix C]. The incentive to include the regularisation method of ridge and elastic net regression is to reduce model complexity. This is significant in obtaining a balance between interpretability and predictability for our models. Using cross validation, we can optimise the regression by performing variable selection using cross validation to achieve the optimal shrinkage parameter. From the results we observed that elastic net gives us the impetus to remove the Downtown variable, whereas none of the covariates approach 0 in ridge regression. In hindsight, it is ideal to prioritise the elastic net method, because ridge regression does not remove any of the covariates. However, the cross validation MSE results showcase that the ridge regression performs much better compared to LASSO and Elastic Net.

KNN models are simple to implement and can handle non-linearities well. As a nonparametric model, fitting the model is also less expensive on average, as it does not require computations of particular parameters or values. Additionally, nonparametric models require fewer assumptions to be considered and implemented, unlike OLS estimations. More

importantly, we are able to use cross-validation to determine the optimal K value, which is a highly effective method of assessing models and preventing cases of overfitting the training data.

However, KNN methods are slower during the prediction process as it is required to search through all the training data points in order to find the ones in the closest proximity to the query point. We observed this reduced speed during the prediction task for the KNN model, especially during the trial and error process of selecting different feature subsets which would outperform and yield a lower prediction error than that of the LASSO-selected covariate subset. The slowness of this method can also be attributed to the relatively large size of the training dataset. Another limitation of this KNN regression model is its lack of interpretability. Unlike with the OLS linear regressions such as the SLR and MLR models, we are unable to quantify the effects that each predictor has on *Withdraw* nor identify the covariate which has the most significant or largest impact on it. A consequence of this is that we are also unable to use KNN methods for feature selection. Further possibilities for KNN methods with our given data is applying KNN classification rather than KNN regression. As we observe multimodality in the dependent variable *Withdraw* [Appendix A], we can assume that there are three commonly withdrawn cash amounts as discussed in the EDA section. Using KNN, we can predict which of the three groups the query data point belongs to, which might yield higher prediction performance than KNN regression models.

There were high expectations for the predictive performance of the deep learning model. Unlike other traditional machine learning algorithms which reach a certain level where their performances cease to improve, neural networks continually improve with more training data. Whilst this is an advantage, neural networks do require more data than other algorithms to have optimal performance, and as a result, is more computationally expensive than other models. However, this did not pose a big concern as our data was fitted optimally with minimal layers. Alternatively, it is their “black box” nature which renders neural networks difficult to work with, as we were unable to attribute changes in the output to changes in the model specifications. As a result, it was hard to fine-tune the model to yield better results without trial and error.

Conclusion

The objective of this report was to develop a predictive model for ATM cash demand. Using the given ATM training dataset, we constructed and fitted different models to assess which model and which specifications of this model would yield the lowest prediction error or MSE when forecasting the total cash amount withdrawn per day from an ATM. Models which were considered and fitted against the data include linear regression models such as OLS, LASSO, Ridge, and elastic net regression models which aided in the feature selection process despite not having the lowest prediction error; kNN regression; and deep learning models. Cross validation allowed for effective assessments of some models, whilst fitting against a validation dataset was used to test the predictive capabilities of other models. Ultimately, we identified the deep learning model to be the optimal model for predicting ATM cash demand, with the lowest CV MSE of 0.2550. As such, we can also expect this model to have the lowest prediction error when we evaluate it with a test dataset. Thus, it is highly recommended that the bank utilises this predictive model to make smarter decisions about reloading its ATM network in order to optimise its cash management.

Reference

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2017. *An introduction to statistical learning*. London.

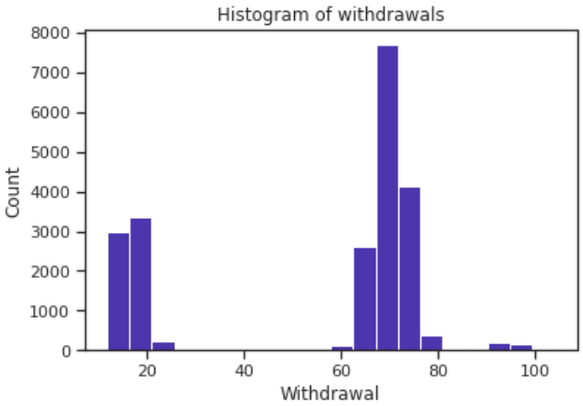
Visionet. 2021. *Optimizing ATM cash management using machine learning - Visionet*. [online] Available at: <https://www.visionet.com/blog/optimizing-atm-cash-management-using-machine-learning/> [Accessed 16 October 2021].

Z. Moubariki, L. Beljadid, M. El Haj Tirari, M. Kaicer and R. O. H. Thami, "Enhancing cash management using machine learning," *2019 1st International Conference on Smart Systems and Data Science (ICSSD)*, 2019, pp. 1-6, doi: 10.1109/ICSSD47982.2019.9002731.

Appendix

Appendix A - Exploratory Data Analysis (EDA)

Histogram of Withdraw



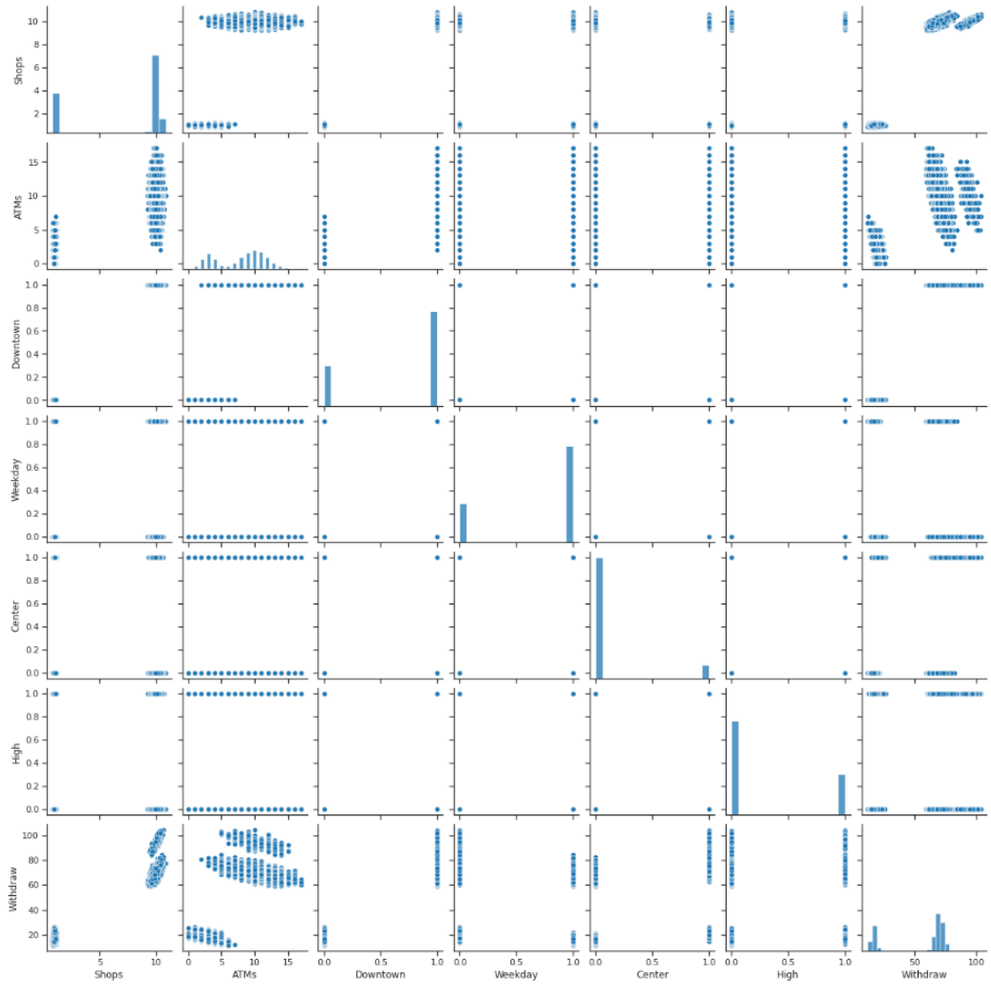
All the covariates that will be used in our data analysis

	Shops	ATMs	Downtown	Weekday	Center	High	Withdraw
0	10.18	10	1	0	0	0	72.750556
1	9.74	10	1	1	0	0	66.720482
2	0.96	2	0	0	0	1	19.189516
3	9.58	9	1	1	0	1	67.388669
4	1.03	4	0	1	0	1	15.813127

Basic summary statistics of the covariates in our data analysis

	Shops	ATMs	Downtown	Weekday	Center	High	Withdraw
count	22000.000000	22000.000000	22000.00000	22000.000000	22000.000000	22000.000000	22000.000000
mean	7.316373	7.937455	0.70200	0.714091	0.102455	0.301591	54.652818
std	4.118692	3.673415	0.45739	0.451857	0.303252	0.458959	25.099767
min	0.800000	0.000000	0.00000	0.000000	0.000000	0.000000	11.668197
25%	1.050000	4.000000	0.00000	0.000000	0.000000	0.000000	18.500386
50%	9.890000	9.000000	1.00000	1.000000	0.000000	0.000000	68.240749
75%	10.070000	11.000000	1.00000	1.000000	0.000000	1.000000	71.345778
max	10.830000	17.000000	1.00000	1.000000	1.000000	1.000000	103.964065

Pairplot displaying distribution between covariates



Appendix B - Simple and Multiple Linear Regression results

OLS Regression of independent variable Shops

OLS Regression Results

Dep. Variable:	Withdraw	R-squared:	0.972
Model:	OLS	Adj. R-squared:	0.972
Method:	Least Squares	F-statistic:	7.580e+05
Date:	Tue, 09 Nov 2021	Prob (F-statistic):	0.00
Time:	00:51:56	Log-Likelihood:	-62868.
No. Observations:	22000	AIC:	1.257e+05
Df Residuals:	21998	BIC:	1.258e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	54.6528	0.028	1923.056	0.000	54.597	54.709
Shops	24.7433	0.028	870.616	0.000	24.688	24.799

Omnibus:	17120.298	Durbin-Watson:	1.985
Prob(Omnibus):	0.000	Jarque-Bera (JB):	381337.712
Skew:	3.649	Prob(JB):	0.00
Kurtosis:	22.046	Cond. No.	1.00

OLS Regression of independent variable ATM

OLS Regression Results

Dep. Variable:	Withdraw	R-squared:	0.679
Model:	OLS	Adj. R-squared:	0.679
Method:	Least Squares	F-statistic:	4.654e+04
Date:	Tue, 09 Nov 2021	Prob (F-statistic):	0.00
Time:	01:07:48	Log-Likelihood:	-89619.
No. Observations:	22000	AIC:	1.792e+05
Df Residuals:	21998	BIC:	1.793e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9.9614	0.228	43.637	0.000	9.514	10.409
ATMs	5.6304	0.026	215.724	0.000	5.579	5.682

Omnibus:	832.081	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	928.269
Skew:	0.498	Prob(JB):	2.68e-202
Kurtosis:	3.150	Cond. No.	21.0

OLS Regression of multiple covariates

OLS Regression Results

Dep. Variable:	Withdraw	R-squared:	0.990
Model:	OLS	Adj. R-squared:	0.990
Method:	Least Squares	F-statistic:	3.656e+05
Date:	Wed, 10 Nov 2021	Prob (F-statistic):	0.00
Time:	00:49:23	Log-Likelihood:	-51380.
No. Observations:	22000	AIC:	1.028e+05
Df Residuals:	21993	BIC:	1.028e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	10.4284	0.111	94.198	0.000	10.211	10.645
Shops	10.8138	0.098	110.062	0.000	10.621	11.006
ATMs	-1.0096	0.009	-106.982	0.000	-1.028	-0.991
Downtown	-36.1897	0.887	-40.781	0.000	-37.929	-34.450
Weekday	-3.5011	0.037	-93.806	0.000	-3.574	-3.428
Center	7.1931	0.056	129.352	0.000	7.084	7.302
High	0.9566	0.037	26.035	0.000	0.885	1.029

Omnibus:	17745.344	Durbin-Watson:	1.998
Prob(Omnibus):	0.000	Jarque-Bera (JB):	468940.833
Skew:	3.781	Prob(JB):	0.00
Kurtosis:	24.316	Cond. No.	646.

Appendix C - LASSO, Ridge and Elastic Net Regression

LASSO Regression coefficients with shrinkage parameter $\lambda = 1$

	Shops	ATMs	Downtown	Weekday	Center	High
0	23.7506	-0.0	0.0	-0.5672	1.2072	0.0

LASSO Regression coefficients with shrinkage parameter $\lambda = 10$

	Shops	ATMs	Downtown	Weekday	Center	High
0	14.7428	0.0	0.0	-0.0	0.0	0.0

LASSO Regression coefficients with optimal shrinkage parameter

	Shops	ATMs	Downtown	Weekday	Center	High
0	27.9023	-3.6249	-0.0	-1.5515	2.1591	0.4157

Ridge Regression coefficients for covariates

	Shops	ATMs	Downtown	Weekday	Center	High
0	44.5388	-3.7088	-16.5527	-1.582	2.1813	0.439

Elastic Net Regression coefficients for covariates

	Shops	ATMs	Downtown	Weekday	Center	High
0	27.8697	-3.5956	0.0	-1.551	2.1586	0.4158