Cash Management Optimisation for ATM Network

James Phan, Nancy Le, Lily Edwards, Hye Jun Jee, Wanda Kuai

10th November 2024

Introduction

The aim of this report is to provide a comprehensive model that informs and aids in cash management optimisation for banks, specifically concerning ATM networks. As economies shift towards a cashless society, prediction of cash demand from ATMs becomes increasingly difficult. Within this report, multiple modelling techniques will be explored and rigorously tested to provide the most accurate method for prediction of cash demand for any given ATM in the bank's network.

Understanding cash withdrawal patterns and the drivers behind ATM usage is crucial to maximising operational efficiency, ensuring satisfaction levels of customers remain high, and safeguarding the bank's liquidity. Utilising the covariates supplied - Shops, ATMs, Downtown, Workday, Center, High - will give insight into the time and location based influences on cash demand within the ATM network being considered.

ATMs with excess cash will have lowered probability of running out of funds and hence will avoid customer dissatisfaction, but will also incur higher costs associated with security, also reducing the amount of cash the bank has on hand for non-ATM uses that generate higher revenue, such as lending. Conversely, too little cash in an ATM will lead to a higher likelihood of running out of funds, increase replenishment costs, but will also allow the bank to use the retained cash for other uses. Minimising the probabilities of the scenarios mentioned can be done through predicting the 'correct' amount of cash for a given ATM, hence why this report focuses on prediction power rather than interpretability.

Through exploratory analysis, we found that cash demand was heavily influenced by specific time and location based factors, leading to clustering within the dataset. This clustering led to the consideration of best subset selection on each cluster of the dataset to aid in prediction performance, along with regularisation and polynomial regression.

Each modelling technique was rigourously assessed in predictive accuracy, but the final model selected for the bank's cash optimisation needs, was the clustered Ordinary Least Squares (OLS) model. This model utilises either best subset selection or forward stepwise selection to aid in model specification, and performs OLS within each cluster of the data to account for different relationships between the factor terms. Clustered OLS outperforms the other models considered in this analysis when considering cross validation test mean-squared-error (MSE) as our measure of predictive power.

Exploratory Data Analysis

Initial Thoughts

In examining the variables provided, a few anticipated trends emerge regarding daily cash withdrawals, the response variable. As briefly described in Table 1, several covariates are expected to have a notable impact on withdrawal levels. For instance, the presence of shops and restaurants nearby (Shops) may correlate with higher withdrawal rates, as such locations generally increase foot traffic and attract individuals likely to make cash transactions. Similarly, a greater number of nearby ATMs (ATMs) might suggest competition, potentially lowering the withdrawal amounts at each individual machine. We also expect contextual factors such as Downtown and Center locations to influence withdrawal amounts. ATMs situated downtown or in central locations (e.g., shopping centers or airports) may experience higher withdrawals due to increased demand in busy or high-traffic areas. Additionally, Workday is anticipated to impact withdrawal levels, with higher withdrawals likely during holiday periods compared to regular workdays, as people may access cash more frequently for leisure activities, shopping, or travel during holidays. Finally, ATMs labeled as High demand in the previous month may indicate

a trend where users rely on these locations, further increasing expected withdrawals. These initial trends offer a foundation for exploring potential relationships between the response and predictor variables, helping guide subsequent data analysis.

Variable	**Description**
Withdraw	The total cash withdrawn a day (in 1,000th of local currency units)
Shops	Number of shops/restaurants within a walkable distance
ATMs	Number of other ATMs within a walkable distance
Downtown	=1 if the ATM is in downtown, 0 if not
Workday	=1 if the day is workday, 0 if holiday
Center	=1 if the ATM is located in a center (shopping, airport, etc.), 0 if not
High	=1 if the ATM had a high cash demand in the last month, 0 if not

Table 1: Description of Variables

OLS Regression

We performed an initial OLS regression to explore the relationships between Withdraw and other predictors. This analysis can be seen as a baseline for understanding each predictor's contribution to withdrawal amounts and might reveal potential multicollinearity.

We will now outline our key findings from conducting an OLS regression. The model achieves an R-squared of 0.990, indicating that the predictors effectively explain 99% of the variance in Withdraw. This high R-squared suggests a strong predictive impact but may also reflect multicollinearity.

The coefficient of all predictors also provides detailed insight into how each predictor influences Withdraw:

- Shops (0.1081): This positive coefficient suggests that areas with more shops see higher ATM withdrawals, likely due to increased foot traffic.
- ATMs (-1.0096): The coefficient for ATMs is negative, implying that locations with more ATMs may see slightly reduced withdrawals. This may reflect dispersed demand, as users have more ATM options in these areas, potentially lowering the volume per ATM.
- Downtown (-36.1897): The large negative coefficient for Downtown is unexpected and could be a result of complex interactions with other predictors like Shops. This finding, though statistically significant ($p < 0.001$), suggests that multicollinearity may be affecting the interpretation of Downtown's impact.
- Workday (-3.5011): This indicates slightly lower withdrawals on workdays compared to holidays, potentially reflect typical patterns in cash usage where people withdraw more cash for holiday activities.
- Center (7.1931) and High (0.9566): These positive results support the idea that ATMs in central or high-traffic areas experience higher withdrawals. These predictors are statistically significant, indicating that ATMs located in central or high-traffic areas play a role in determining withdrawal.

Additionally, a high conditional number (4.45e+04) suggests strong multicollinearity, particularly among Shops, ATMs and Downtown which could lead to instability in the coefficient estimates. This observation indicates that coefficients for some predictors may be unstable due to correlations among the features.

```
==============================================================================
                            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:                Sun, 10 Nov 2024   Prob (F-statistic):               0.00
Time:                        09:43:12   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          0.1081      0.001    110.062      0.000       0.106       0.110
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
Workday       -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.                     4.45e+04
================================================================================
```

Multicollinearity

Scatterplots and Heatmaps

Overall Dataset

Initial scatterplots of Withdrawals vs. Number of Shops and Withdrawals vs. Number of ATMs reveal three distinct clusters within the dataset, as shown in Figure 1. These clusters indicate that withdrawal behaviour may vary across specific subsets of the data, potentially due to geographic factors.

In the Withdrawals vs. Shops scatterplot, a strong positive correlation emerges, aligning with our hypothesis that ATMs located near a higher density of shops tend to experience increased cash withdrawals due to greater spending opportunities. However, cluster 3 deviates from this trend, showing a weaker correlation. This suggests that additional factors may be influencing withdrawal behavior within this subset, indicating that a more sophisticated model might be needed to fully capture these dynamics.

Conversely, in the Withdrawals vs. ATMs scatterplot, a consistent strong negative linear correlation is observed, which fits with our initial expectations. Here, as the number of nearby ATMs increases, the demand for cash withdrawals at individual ATMs decreases, as users have more options available, spreading the total cash demand across multiple ATMs.

Additionally, a heatmap of the entire dataset revealed multicollinearity between certain location-related variables—particularly between the downtown variable and the counts of both shops and ATMs. This multicollinearity suggests that the downtown variable may indirectly influence withdrawal rates through its associations with nearby amenities. We also observed an unexpected finding in the heatmap: Withdrawals and number of ATMs show a positive correlation at the dataset level, which we later see, contrasts with the negative correlation observed in individual clusters. This positive correlation may indicate that, at an aggregate level, areas with higher ATM density also have increased overall cash demand, even though individual ATM withdrawals decrease due to increased competition.

Overall, this insight further provokes the analysis of individual clusters. It suggests that viewing the dataset as a whole may mask important distinctions within clusters where different patterns, potentially influenced by localised factors, are at play. The contrasting trends between the scatterplots and the heatmap underscore the need to examine each cluster individually to understand how factors like location, density of shops, and ATM availability interact to drive withdrawal behaviours in specific areas. By examining each cluster more closely, we can better isolate the regional factors that impact ATM usage and potentially refine our model to capture these localised effects more accurately.

Individual Clusters

To further investigate the distinct withdrawal patterns observed in our initial analysis, we proceeded by splitting the dataset into two categories: downtown and not downtown. This separation effectively distinguishes the lower cluster (representing areas not classified as downtown) from the upper two clusters, which correspond to downtown locations, in both scatterplot graphs. Notably, this segmentation resolves the earlier issue of a positive correlation between Withdrawals and Number of ATMs. In the updated heatmaps for both downtown and not downtown, we observe the expected negative correlation, confirming our hypothesis that increased ATM density typically leads to lower individual ATM withdrawals as cash demand is more distributed. Furthermore, by isolating the dataset into these two categories and thereby excluding the downtown variable as a predictor, we eliminate multicollinearity issues among the remaining predictor variables.

When examining the not downtown category, the scatterplot comparing cash withdrawn with the number of shops reveals a more ambiguous relationship. Unlike the clear patterns observed in the downtown clusters, there is no evident linear correlation in the not downtown group. This lack of a discernible relationship is further substantiated by a correlation coefficient of just 0.035 in the corresponding heatmap, indicating a very weak association. Consequently, this ambiguity suggests that a more complex model may be necessary to adequately capture the dynamics at play in non-downtown areas.

With our analysis of the downtown and not downtown segments underway, we are left with two clusters that require further distinction. To achieve this, we seek to identify a combination of categorical variables that effectively classifies these remaining clusters. The uppermost cluster can be categorised as downtown + holiday + center, indicating that withdrawal behaviours in this cluster are influenced by a concentration of shops, higher foot traffic during holidays, and a central location. The other cluster can be characterised as a combination of downtown + weekday and downtown + holiday + not center. This classification reflects a more complex interaction between location and time, where withdrawals differ based on whether it is a workday or a holiday, as well as the center's accessibility.

The intuitive appeal of incorporating holiday and center as distinguishing factors is clear. Holidays often lead to increased consumer spending and foot traffic, which could drive higher withdrawal volumes. However, the question arises: how does the combination of these two variables justify the existence of a separate and higher withdrawal cluster compared to the others? This distinction highlights the need for deeper exploration into how these factors interact to influence withdrawal patterns.

It is worth noting that attempts to separate the remaining two clusters by splitting the dataset into center/not center or holiday/not holiday categories were unsuccessful. These approaches did not yield distinct separations, indicating that a more nuanced understanding of the factors at play is required.

Unlike the ambiguity observed in the not downtown dataset, there is a clear linear relationship between withdrawals and the number of shops and ATMs in the remaining clusters. This linearity suggests that, within these specific downtown contexts, as the density of shops and ATMs increases, so too does the cash withdrawn. This pattern reinforces the necessity of considering both temporal and spatial factors in our analysis to accurately depict the dynamics of ATM withdrawals.

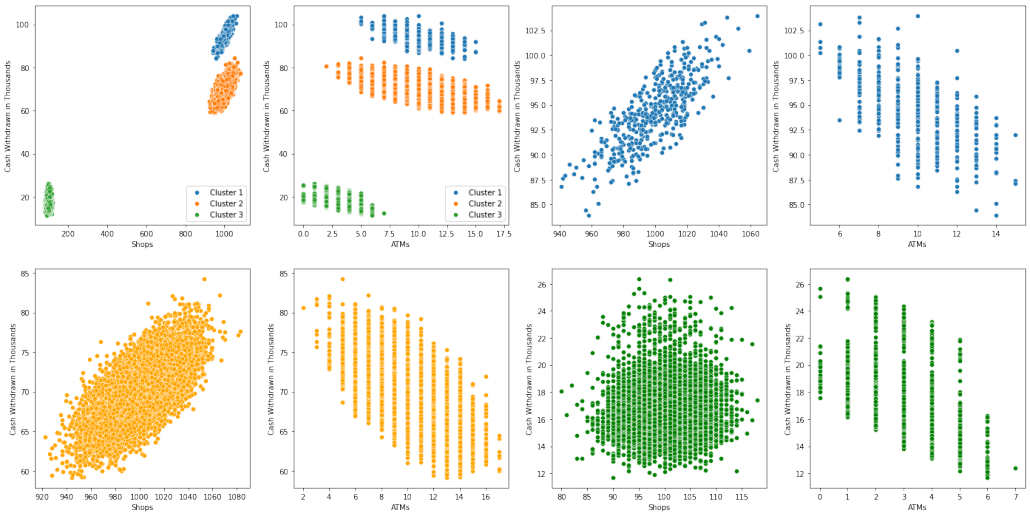


Figure 1: Scatter plots illustrating relationship between cash withdrawals with Shops, and ATMs across three distinct clusters. Cluster 1 includes areas that are downtown, have holiday activity, and are located in central regions. Cluster 2 consists of downtown areas with holiday activity or workday patterns, and may or may not be located in central areas. Cluster 3 represents areas that are not downtown.

Figure 2: Heat map with correlation coefficients amongst the variables on the entire dataset, as well as individual clusters

Correlation Matrix and VIF

It is ideal to know not only whether there is multicollinearity in the model, but also what degree of problem we have (weak, moderate, strong, etc) and determine which predictor variable(s) causes the problem. Therefore, to detect multicollinearity following methods can be used:

Scatter Plot/correlation matrix (Visualised by heatmap)	Versatile but it can only reveal near-linear dependence between a pair of predictors
Variance inflation factors (VIFs)	Can be used to detect near-linear dependence among any number of predictors
Condition number of the correlation matrix	A large value of condition number κ ($= \lambda_{\max} / \lambda_{\min}$) > 1000 indicates strong multicollinearity

As the scatterplot matrix of all predictors can be considered a subjective way for the detection as we've seen above, computing the pairwise correlation score is a better approach. From the heatmap above, we can confirm the following regressor pairs are highly correlated: (Shops, Downtown), (Shops, ATMs), (Downtown, ATMs) and (Shops, Withdraw), (Downtown, Withdraw) being highly correlated with the target variable "Withdraw".Furthermore there are few non-linear relationships such as (Shops, Withdraw) as well. The correlation matrix of the predictors are often used to detect collinearity. However, unfortunately, not all collinearity problems can be detected by correlation matrix as the collinearity can exist between three(or more) variables, instead of pair ones, even if no pair of variables has a particularly high correlation, and this kind of situation is called multicollinearity.

To assess multicollinearity, computing the variance inflation factor(VIF) is considered better than inspecting the correlation matrix (Gareth et al., 2023, p.108). The formula of the VIF for each variable is

$$VIF(\hat{\beta}_j) = \frac{1}{1 - {R^2}_{X_j|X_{-j}}}$$

Where ${R^2}_{X_j|X_{-j}}$ is the R^2 from the regression of X_j onto all the other predictors. The minimum possible value of VIF value is 1, indicates no correlation and the value above 10 indicates severe multicollinearity.

	Variable	VIF
0	const	44.180886
1	Shops	498.645266
2	ATMs	3.826736
3	Downtown	501.806147
4	Workday	1.032032
5	Center	1.003322
6	High	1.011151

(a) VIF before removing "Downtown"

	Variable	VIF
0	const	11.941903
1	Shops	3.825804
2	ATMs	3.803325
3	Workday	1.031792
4	Center	1.003306
5	High	1.011136

(b) VIF after removing "Downtown"

Figure 3: Comparison of VIF values before and after removing the "Downtown" variable.

The VIF values on Shops decreased significantly after removing the Downtown variable and it implies that the strong correlation of (Shops, Downtown) was a key contributor to multicollinearity, justifying the removal of the variable "Downtown". The VIF analysis on the dataset allowed us to identify and address the source of multicollinearity, but dealing with multicollinearity calls the need to do Ridge regression (and Lasso) and use model selection methods to eliminate the redundant predictors from the model.

Model Analysis: Models and Methods

Regularisation

Based on the initial OLS analysis, we observed significant multicollinearity among several predictors, particularly Shops, Downtown, and ATMs. Additionally, the correlation matrix also highlights these relationships, with a perfect correlation between Shops and Downtown and a high correlation between Shops and ATMs. This multicollinearity can lead to unstable coefficient estimates in standard linear regression models. To combat

multicollinearity, regularisation introduces a penalty term to all model parameters, (excluding the intercept) to limit the size of the coefficients of correlated covariates, which is an effect of multicollinearity. We explore three regularisation techniques: ridge, lasso and elastic net. A comparison of these techniques is discussed below.

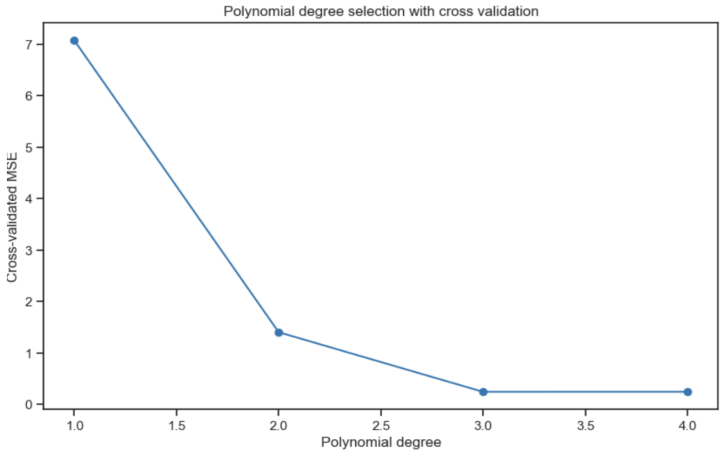
Regularisation technique	**Penalty term**	**What does each regularisation do?**
Ridge	$\lambda \sum_{j=1}^{p} \beta_j^2$	• Coefficients may be shrunk towards zero, but not to zero • Reduces variance of parameter estimates although bias is introduced • Resolves multicollinearity by limiting the size of the estimated coefficients of predictors
Lasso	$\lambda \sum_{j=1}^{p} \lvert\beta_j\rvert$	• Coefficients of less significant predictors may be shrunk either to zero or towards zero • Creates sparsity in the model and can be used for variable selection
Elastic net	$\lambda \sum_{j=1}^{p} \left(\alpha\beta_j^2 + (1-\alpha)\lvert\beta_j\rvert\right)$	• Combines ridge and lasso penalty terms • Effective in correcting multicollinearity and variable selection simultaneously

Table 2: Comparison of ridge, lasso, and elastic net regularisation techniques.

We implemented each of the above regularisation techniques by experimenting with polynomial and interaction features, splitting the dataset into training and test sets, standardising inputs before fitting the model, and then testing model performance by calculating the CV MSE on the test dataset. Firstly, we would find the MSE for the base model (i.e. without interactions and covariate terms have degree 1). To explore methods to reduce the MSE, we introduced squared and cubic terms of the continuous covariates, Shops and ATMs, and interaction terms between these polynomial terms and the given covariates to capture any non-linear relationships. In the case of lasso, all polynomial and interaction terms up to degree 3 when lasso was used with k-means clustering. Across all three regularisation methods, we observed that adding these terms into our model resulted in a lower MSE.

To implement the regularisation models with cross-validation to find the MSE for an optimal alpha, the main library we used was Scikit-Learn. In particular, the main functions used consisted of: train_test_split, Kfold, Ridge, Lasso, ElasticNetCV, ElasticNet, and cross_val_score. Prior to any model training or testing, the covariates were standardised using StandardScaler from Scikit-Learn, to ensure proportionate penalties, this is due to the fact that the penalty value is dependent on the magnitude of the coefficient, hence non-standardized data will lead to highly variable penalties and more biased estimates. Across our regularisation models, the data was split using train_test_split with a test size of 20 percent using a random state of 42 to ensure reproducibility and consistency. The reasoning behind a training and test split is to aid in the prediction capabilities given an independent dataset, it gives a measure of quality for the ultimate model (Hastie, Tibshirani and Friedman, 2009, p.219).

Building on the above regularization foundation, we aimed to enhance predictive accuracy by exploring the optimal polynomial degree as follows. The following figure shows the cross-validation results for various polynomial degrees in a graph, aiding in the selection of the optimal degree by providing information needed to minimize potential overfitting issues and preserve model generalizability.



Cross-validation for selecting optimal degree results shows 7.0747, 1.4062, 0.2479, and 0.2483 on degrees 1 to 4. This shows that the degree 3 polynomial is optimal as the cross-validated MSE (CV MSE) gradually decreases till degree 3 and it slightly increases on degree 4, implying any degree greater than 3 might giving the model overfitting.
For further cross-validation for degree 3 polynomial regression, we use KFold from scikit-learn library to do residual analysis to evaluate the model's fit. It aims to find any possible improvements of the degree 3 polynomial model. The calculation of CV MSE over folds are as following:

```
Cross-Validation MSE Scores: [0.25545065 0.25243451 0.24438755 0.2423539
0.24400505]
Mean Cross-Validation MSE: 0.2477
Mean of residuals: 0.0000
Standard deviation of residuals: 0.4970
```

Figure 4: CV MSE scores and residual results for a degree 3 polynomial regression

Ridge

The penalty term in ridge regression shrinks coefficients of correlated predictors, thereby improving model stability and interpretability. To identify the optimal model configuration, we evaluated several setups, varying polynomial degree and the inclusion of interaction terms. The configurations tested included:

- Base Features Degree 1 with and without interaction terms
- Polynomial Degree 2 with and without interaction terms
- Polynomial Degree 3 with and without interaction terms

The model testing showed a clear advantage for configurations that included interaction terms. Polynomial Degree 3 with Interactions achieved a validation MSE of 0.2510, a substantial reduction compared to the MSE of 6.1275 for the same degree without interactions. Similar trends were observed with Polynomial Degree 2, where including interactions lowered the validation MSE to 1.6410 compared to 6.1271 without interactions. These results indicate that interaction terms capture complex, meaningful relationships between predictors, significantly enhancing the model's predictive performance.
To train and validate each configuration, we followed:

- Polynomial Feature Expansion: For models with interactions, PolynomialFeatures was used to generate all polynomial and interaction terms up to degree 3.
- Standardization: Features were standardized using StandardScaler to ensure that regularization applied uniformly across predictors.
- Cross-Validation: A range of alpha values, from 10^{-4} to 10^{4}, was tested using 5-fold cross-validation. This helped identify the optimal regularisation strength, with alpha values chosen based on the model configuration that minimised MSE.

Based on validation MSE, the model with Polynomial Degree 3 with Interactions and the best alpha value of 0.0193 was selected as the selected model configuration.
The chosen model achieved an R-squared of 0.9996, indicating it explains nearly all the variance in ATM withdrawals. The MSE was 0.2474 and cross validation test MSE was 0.2487, demonstrating high predictive accuracy and a significant improvement over simpler models without interactions. After applying regularization, there is some key coefficient that influences Withdraw:

- $Shops^2$ (6.7877) and $Shops^3$ (4.4452): These quadratic and cubic terms reveal a non-linear relationship between shop density and ATM withdrawals. The diminishing returns suggest that while higher shop density initially boosts withdrawals, the impact tapers off as density increases.
- $Shops^2 Downtown$ (6.4477): This interaction term highlights that the effect of shop density is particularly pronounced in downtown areas, where ATM withdrawals are significantly higher. This finding suggests that downtown locations offer greater potential for ATM usage due to concentrated commercial activity.
- $ATMs$ (-3.7028): The negative coefficient for ATMs indicates that adding more ATMs in a single area leads to slightly lower withdrawals per ATM, reflecting a distributed demand across multiple machines.
- $ShopsWorkdayCenter$(-4.7680) and $ShopsATMsDowntown$(1.2127): These interaction terms underline the varying impact of locational and temporal factors on withdrawal behavior. For example, the negative coefficient for Shops Workday Center suggests that the effect of shop density is moderated during weekdays in central locations, possibly due to different customer behaviors on workdays.

```
Model Summary:
==================================
Best Alpha: 0.019306977288832496
R-squared: 0.9996
Mean Squared Error (MSE): 0.2474
AIC: -30613.3491
BIC: -30165.4164
==================================
```

Feature	Coefficient	Std Error
Intercept	54.652818	NaN
Shops^2	6.787656	3.084617
Shops	6.592232	1.211432
Shops^2 Downtown	6.447716	2.948906
Shops Workday Center	-4.768004	0.313408
Shops Downtown	4.648742	2.851448
Shops^3	4.445164	1.796655
Shops Center	4.000660	1.097849
ATMs	-3.702765	0.230480
Shops^2 ATMs	-1.293415	2.053351
Shops ATMs Workday	1.236881	0.989502
Shops Downtown High	-1.222662	2.755235
Shops ATMs Downtown	1.212712	2.931315
Downtown	-1.168996	1.638568
Center	1.016487	0.125169
ATMs Downtown Workday	-0.995535	0.909528
Workday	-0.881871	0.116670
Shops^2 Center	0.860075	1.577282
Shops^2 Workday	-0.759804	1.607889
Shops^2 High	0.565613	1.580212
Downtown Workday	0.515912	1.610618
Shops ATMs Center	-0.493818	0.789566
Shops Downtown Workday	0.489574	2.755307
Shops High	0.428070	1.141088
High	0.394188	0.125934
ATMs Downtown	0.390484	2.027555
Shops Downtown Center	0.381943	2.752699
Shops Workday	-0.345567	1.276353
Downtown Workday Center	0.333025	0.283238
Shops ATMs	-0.317027	2.724996
ATMs Downtown Center	0.309256	0.717179
ATMs^2 Downtown	-0.295188	1.312240
Downtown High	0.270401	1.588898
Shops ATMs High	0.264714	0.838526
ATMs Downtown High	-0.200603	0.765378
Shops Workday High	-0.174004	0.333520
Shops ATMs^2	0.157486	1.411799
ATMs^2	0.136527	0.294017
Downtown Center High	0.125778	0.188009
Downtown Center	-0.121280	1.573964
Downtown Workday High	0.113563	0.303056
ATMs^2 Workday	-0.108792	0.089198
Shops Center High	-0.105715	0.208224
ATMs^2 High	-0.104564	0.069601
ATMs^2 Center	0.087327	0.064513
Workday Center	0.055900	0.040212
ATMs Workday High	0.054741	0.032883
ATMs Workday	-0.050279	0.119489
ATMs Center	0.042823	0.105577
ATMs High	0.041996	0.109962
ATMs Center High	-0.033286	0.020712
ATMs^3	0.024745	0.152831
Center High	0.016727	0.027468
ATMs Workday Center	-0.015441	0.031410
Workday High	0.007913	0.041958
Workday Center High	0.000495	0.007872

Lasso

Contrary to ridge regularisation, the hyperparamater in lasso's penalty term, λ, is capable of automatically selecting significant covariates. As the value of λ increases, more coefficients of non-significant covariates are shrunk towards zero, if not zero. Therefore, if there is multicollinearity among covariates, lasso's feature selection will only keep the most influential covariates and reduce the influence of other correlated covariates.

Preliminary Testing
The MSE for the base lasso model, which was built according to the steps described above, is 6.2316. To lower the MSE with a lasso regression, we introduced interaction terms and polynomial terms (i.e. squared and cubic Shops and ATMs) to reduce the MSE and capture any non-linear relationships between Withdraw and the predictors. The MSE was only minimally reduced to 5.6422. Contrary to ridge and elastic net (below), lasso's test MSE obtained with interaction and polynomial terms is relatively high. We suspect that this is due to lasso enforcing sparsity when selecting the most significant predictors.

Lasso Regression and K-Means Clustering

As a result, we opted for an alternative method to reduce the MSE. We opted to conduct a k-means cluster as a preprocessing step before applying a lasso regression model to each cluster. As an additional enhancement, we continued to add in polynomial and interaction terms with PolynomialFeatures with degree set to 3. K-means clustering is a distance-based algorithm that partitions a dataset into a pre-defined number of clusters according to similarity in the data (Hastie, Tibshirani and Friedman, 2009, p.510). Therefore, applying a lasso regression to each cluster ensures we practice targeted modelling to capture cluster-specific trends that may not be observed in a global dataset, as will also be excluded in our clustered OLS regression model. The first step in the implementation was to determine the number of clusters, which was determined by silhouette_scores from sklearn.metrics. This gave the result that the optimal cluster number was 4. In a for loop, we fit a lasso regression model to each of the 4 clusters and stored the test MSE, best_lasso_model and R^2 values for each cluster. To further reduce the MSE, we used GridSearchCV to fine-tune the hyperparameter, which in the case for lasso is the regularisation strength, alpha. The average test MSE of 0.2429, a weighted average test MSE of 0.2468, an average CV MSE of 0.2433 and a weighted average CV MSE of 0.2489. We performed a weighted MSE to ensure the model performance was proportionate to each cluster size, to ultimately improve the stability of estimates and improve the predictive power.

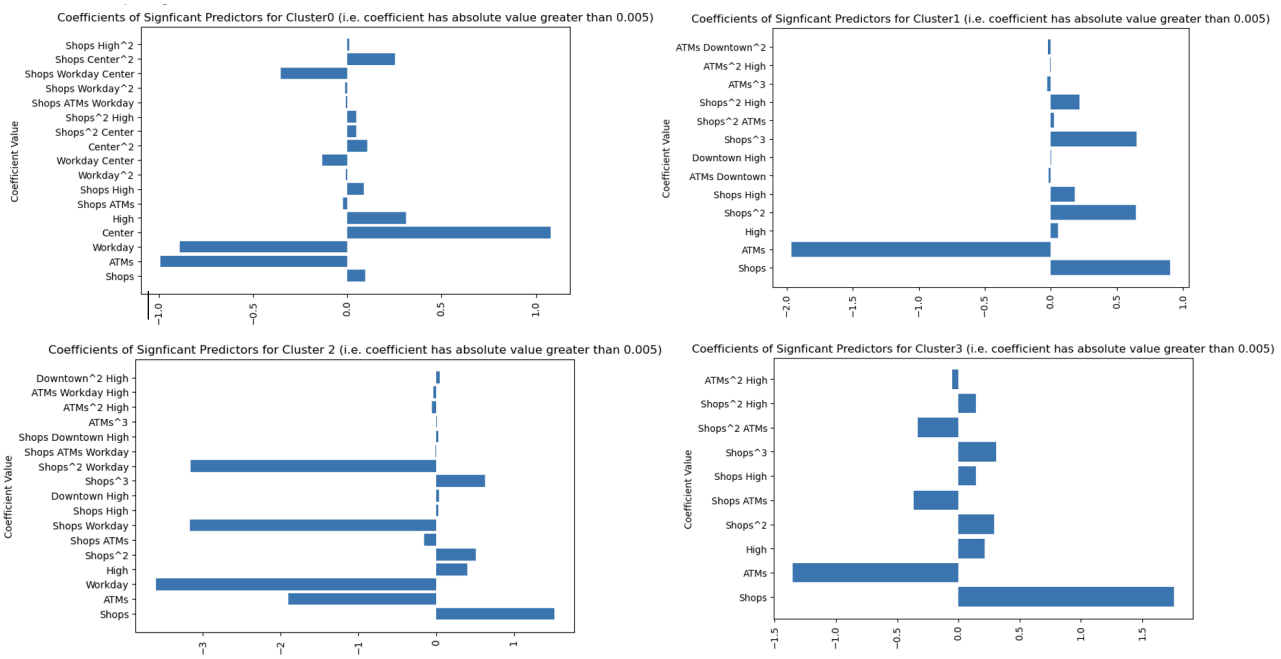


Figure 5: Plot showing the size of the coefficients against the corresponding predictors. To select significant predictors, we excluded coefficients with an absolute value less than 0.005, which includes any predictors with coefficients pushed directly to 0.

As seen in Figure 5, Workday and its interaction terms (including $ShopsWorkdayCenter$, $Shops^2Workday$ and $ShopsWorkday$) have relatively large negative coefficients in Clusters 0 and 2. This observation reinforces our expectations that cash demand is higher during holidays, to the extent that withdrawal demand on holidays is also related to the time of day (i.e. $Workday$) and location (i.e. $Shops$ and $Center$). This suggests that the ATMs in Clusters 0 and 2 should restock on cash before or during holidays to prevent any shortages. Further, cluster 2 has the highest R^2 value of 0.9978 suggesting the best lasso model in cluster 2 has a higher predictive performance relative to other clusters, which reinforces our initial conclusions about the relationships between $Shops$, $ATMs$ and $Withdraw$. Although ATMs is negatively correlated with $Withdraw$ across all four clusters, Cluster 1 has a distinctively high negative coefficient of -1.9637. Therefore, ATMs, in general, is associated with a decrease in demand for each ATM, especially in areas found in Cluster 1. The bank may not need to install additional ATMs in such areas because the existing number of ATMs are capable of satisfying client demand.

It is also unsurprising that cash demand is high when ATMs are in centers or in areas with shops and restaurants, including areas with a high density of shops (i.e. terms with $Shops^2$ and $Shops^3$). Given that this trend was observed across all four clusters, this is strong evidence to support our initial belief that withdrawals is

location dependent and, therefore, higher in locations with higher foot traffic. Further, with the exception of Cluster 0, *Shops* has the highest positive coefficient for the remaining three clusters. These clusters imply cash transactions are preferred for shops and restaurants, which may be due to limited cashless options.

The lasso models across clusters consistently push the coefficient of *Downtown* towards zero, if not zero. As *Downtown* was a key predictor resulting in multicollinearity, lasso's feature selection is consistent with our initial exploratory data analysis as it favours other location variables (including *Shops*, *Centers* and their interaction terms) to predict cash demand. While it is reasonable that a downtown area may result in higher withdrawals, this method indicates that this trend is likely explained by the effect of more significant predictors. Contrary to cash demand in shopping districts, cash demand in downtown areas behaves differently, possibly due to the high availability of digital payment options or consumer preference for cashless transactions in downtown areas. In summary, this clustering-lasso method gives tailored predictions models to improve the bank's management of cash reserve and optimise the bank's operational strategy. Although additional features are added to capture non-linear relationships, the additional use of k-means clustering and lasso may increase computational cost.

Elastic Net

Elastic net is a form of regularisation that acts as a compromise between ridge and lasso regressions, performing both variable selection and shrinking of coefficients to correct multicollinearity issues and prevent overfitting. There are two parameters within the model, one is the l1 ratio, sitting between zero and one, representing the tradeoff between the ridge and LASSO norms, while alpha represents the penalty. For an l1 ratio closer to one, elastic net will behave similarly to LASSO, whilst an l1 ratio close to zero will behave more like ridge (Scikit-learn).

Considering elastic net is a regularisation technique, no covariates were removed even if multicollinearity issues appeared to be present prior to modelling. That is, Downtown and any added interaction terms including Downtown were not removed from the set of covariates due to the nature of regularisation techniques and their inherent ability to manage any multicollinearity issues. Elastic net was performed on the base dataset and an enriched dataset, the enriched dataset includes distinct two-way and three-way interactions, along with quadratic and cubic terms for both continuous variables, Shops and ATMs. New interaction terms were included to capture possible relationships between the variables, and potential non-linearity in Shops and ATMs. The number of interaction terms added could lead to overfitting, but by monitoring the test MSE it will be evident whether this is a legitimate concern, and the elastic net should be able to mitigate any significant issues with overfitting. Considering elastic net utilises both lasso and ridge, it is likely that a significant number of these added interactions will add nothing to the model, and those coefficients will be reduced to zero.

ElasticNetCV was utilised along with a list of potential values for both alpha and the l1 ratio. Cross-validation found the best l1 ratio to be equal to 1, corresponding with lasso, and the best alpha value to be 0.0248. The cross validated parameter values were used with the test set to find the CV test MSE, equal to 0.2809. Performing the same steps on the base model resulted in the very similar values for both the l1 ratio and alpha, but resulted in a cross validated test MSE of 6.6942.

As can be seen in Figure 6 below, the enriched model shrinks a significant amount of covariates to zero, out of the covariates that have non-zero coefficients, Shops along with Shops^2 and Shops^3 seem to have the largest effect on the withdrawal amount. Downtown alone has been removed from the model, likely due to its near perfect collinearity with Shops, but interestingly some interaction terms with Downtown remain, most also include either Shops, Shops^2 or Shops^3, suggesting that the combined effect of Downtown with Shops and other location and temporal indicators has significant effects on Withdraw.

The plot of the base model coefficients in Figure 7, shows that only Downtown was reduced to zero, again, likely due to multicollinearity issues between Shops and Downtown. Evidently prediction improved by including interaction variables in the model, these interaction terms account for unseen relationships between the covariates, and aid in capturing the dynamics between cash demand, location and time variables. Intuitively it makes sense that cash demand is highest in a location with a high density of shops nearby, i.e. a shopping centre. Both the base model and the enriched model result in Shops having the largest coefficient value, suggesting that ATMs in areas with a large number of shops should be allocated more cash than those elsewhere as these ATMs have the largest positive impact on cash demand. It's also evident from both plots that Workday, and any non-zero interactions with Workday that remain have negative coefficients, suggesting cash demand is higher on holidays.

Overall, elastic net is a fairly uncomplicated model when it comes to interpretability, due to it being penalised least squares, the most complicated aspect is finding the optimal parameters, but can easily be done using cross validation.

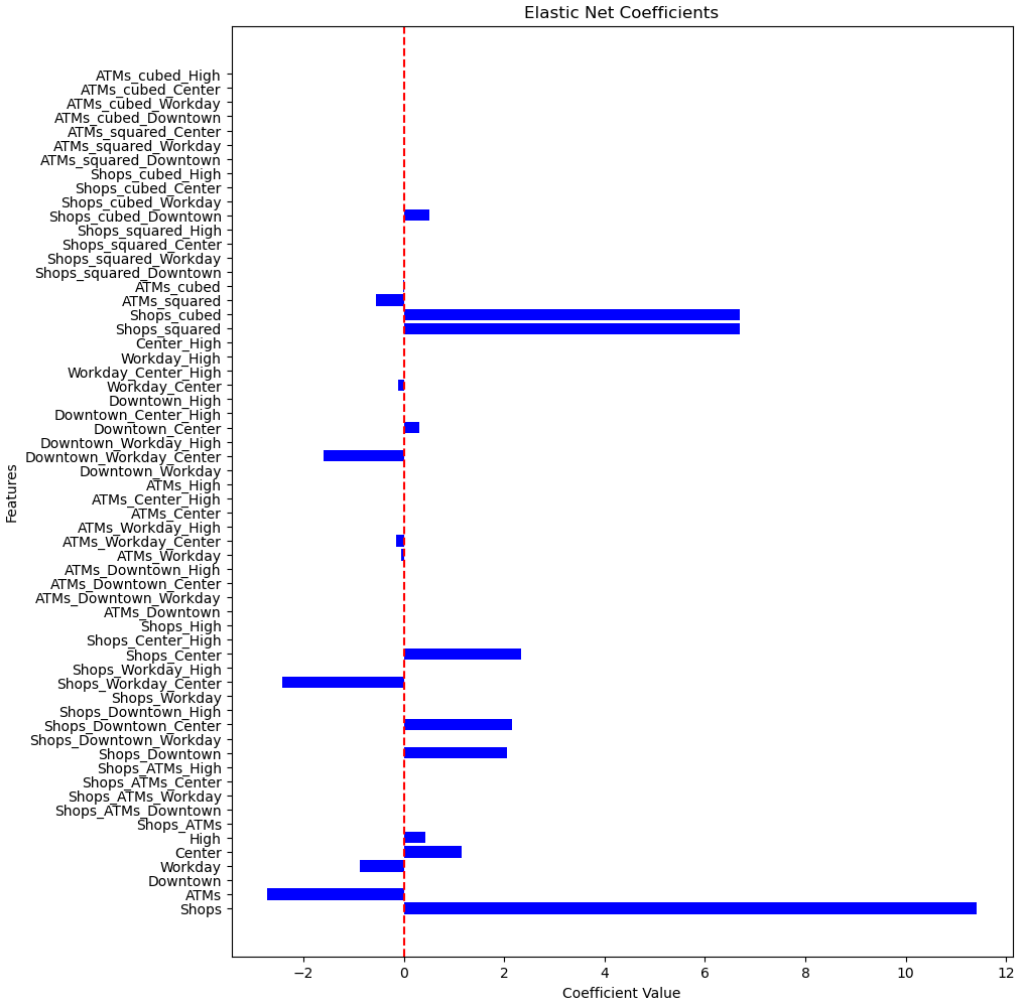


Figure 6

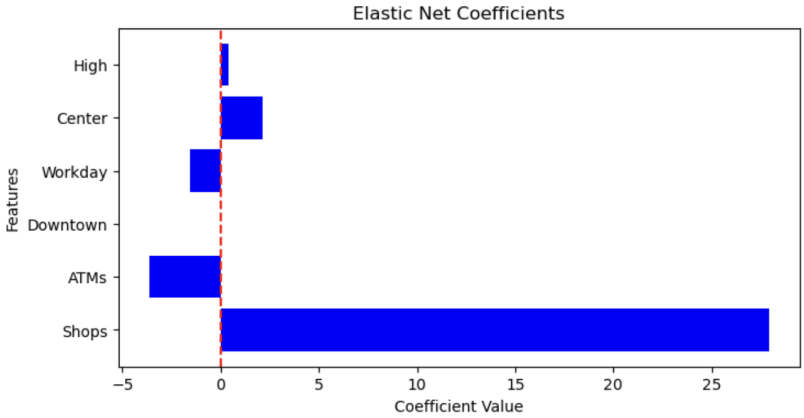


Figure 7

Individual Cluster Regression

Based on our previous cluster analysis, we considered an alternative methodology of developing individual models for each identified cluster, a decision driven by the unique withdrawal behaviors and patterns observed within each.

Justifications for Cluster Regression
The primary aim of this approach is to determine whether a model trained on a specific cluster would outperform previous models trained on the entire dataset. If such an improvement is observed, we could opt to use the cluster-specific model for that particular domain to enhance prediction accuracy for that subset of data and allows for more precise modeling that better captures the specific patterns of withdrawal behavior in each context. Fitting separate models for each cluster also serves to reduce the bias that would arise from using a single model across the entire dataset. While this cluster-specific approach may introduce greater variability (i.e., higher variance) due to the more precise fit to each cluster's data, the trade-off ultimately results in improved predictive performance. By closely aligning each model with the unique data in its segment, we can better account for the detailed distinctions between clusters.

Furthermore, training models exclusively on each cluster theoretically leads to higher accuracy compared to a single, generalized model. Each cluster can be seen as its own distinct domain within the dataset (e.g., downtown + holiday + center versus not downtown). Developing domain-specific models enables us to accommodate the particular influences of location and timing, producing a closer fit and lower bias for each segment. This targeted modeling approach ensures that we account for the specific factors that influence withdrawal behavior within each cluster, which would be overlooked in a whole-dataset model.

A core assumption underpinning this cluster-based approach is that data points can reliably be categorised into one of the clusters based on factors like location, and that these clusters maintain consistent withdrawal patterns within certain ranges. This assumption is crucial, as it supports the relevance of each model to its respective cluster, ensuring accuracy and consistency in predictions.An additional advantage of using cluster-specific regression is its ability to address multicollinearity issues. By splitting the data into distinct clusters and removing location-based indicators within each subset, we reduce the correlation between predictor variables. This reduction in multicollinearity improves model stability and interpretability by isolating the influence of key predictors without interference from highly correlated variables present in the entire dataset.

Implementation
To identify the optimal linear base model for each cluster, we applied OLS regression on all possible base predictors, using Best Subset Selection (BSS) to evaluate all potential model configurations. BSS iteratively considers every possible subset of predictors, fitting an OLS regression for each, and selects the subset that best aligns with our chosen evaluation metric. The OLS regression approach finds coefficients for each predictor by minimizing the sum of squared residuals, allowing us to derive a best-fit line that describes the relationship between predictors and the response variable within each cluster. In our case, we evaluated each model configuration by computing the test MSE on a single validation set. While cross-validation would generally offer a more reliable estimate by reducing variance and mitigating the risk of overfitting, we chose a single validation set for computational efficiency. Future iterations could consider cross-validation to achieve a more robust estimate of test MSE, especially if computational resources allow.

Overall, using OLS with best subset selection was effective for initial modeling, as the resulting models for each cluster demonstrated promising test MSE scores ranging from 0.2138 to 0.3279. However, to further improve model performance, we sought to capture nonlinear relationships and interactions within the data. This was accomplished by creating an enhanced dataset, which included polynomial terms up to the third degree, logarithmic transformations, and two-term interactions between all predictors.

Given the substantial increase in the number of predictors with this enhanced dataset, using BSS was no longer computationally feasible. Instead, we implemented Forward Stepwise Selection (FSS), a more efficient alternative. FSS builds the model by starting with no predictors and sequentially adding the predictor that most significantly reduces the test MSE at each step. While this greedy algorithm does not guarantee finding the global optimal subset, it provides a practical solution when handling large numbers of predictors. Future improvements could involve revisiting best subset selection if computational constraints are relaxed, as it would ensure a truly optimal model configuration.

After applying FSS on the enhanced dataset and running OLS on the selected enhanced features, the only notable improvement is Cluster 3's CV Test MSE, reducing from 0.3279 to 0.2524. This is to be expected, given the nonlinearity of our initial scatter of cluster 3, we'd expect a more complex model which forward stepwise selection has picked up. Cluster 2 only saw a slight improvement using the enhanced features, reducing from 0.2487 to 0.2485, suggesting that the simple linear model holds strong, again as expected from our exploratory analysis. Cluster 1 however saw an increase in model evaluation from 0.2138 to 0.2189. Similar to Cluster 2, we'd expect the best model for Cluster 1 to be approximately linear.

1. Best Subset Selection
BSS involves fitting all possible combinations of predictors and selecting the model with the lowest test MSE. This exhaustive approach ensures that we examine every potential subset of predictors, identifying the combination that best aligns with our chosen criteria.

To implement a BSS, we start with a null model that only includes an intercept (Withdraw ~ 1). This model provides a baseline for comparison against more complex models. Using Python's itertools.combinations, we generate and evaluate every possible subset of predictors, beginning with models of size 1 (single predictor) and incrementing up to the full set. For each subset, we used the statsmodels.ols function to fit an OLS regression model before calculating the MSE on the training set for each model. We track the model with the lowest train MSE for each combination size.After identifying the best model based on train MSE for a specific combination size, we calculate its test MSE using a separate test dataset and a test_MSE function. The test_MSE function computes the MSE on out-of-sample data to evaluate the model's predictive performance on unseen data. If the candidate model's test MSE is lower than the current best test MSE, it becomes the new best model. We conclude BSS by returning the model that achieved the lowest test MSE across all possible predictor combinations.

2. Forward Stepwise Selection
Unlike BSS, FSS incrementally adds predictors one at a time based on their individual contribution to reducing test MSE. This approach is computationally more efficient, as it does not require evaluation of every possible predictor combination. To implement FSS, we begin with a null model (Withdraw ~ 1) and set this as the baseline model. At each step, we consider adding each predictor that is not currently in the model. We evaluate each potential new model's performance on the training set by calculating its train MSE. The predictor that yields the lowest train MSE when added to the model is selected. This predictor is appended to a selected_predictors list. After adding a new predictor, we evaluate the updated model's test MSE. The test_MSE function, as with BSS, calculates the model's predictive performance on the test set. This process continues until no additional predictor improves the model's test MSE, at which point we finalize the model. Forward stepwise selection returns the model with the lowest test MSE, accounting for only those predictors that demonstrated improvement at each addition step.

Model Summaries
Model summaries of the OLS of each cluster are depicted below. Models that are not part of the final product are not listed in this report, but can be seen in our implementation file.

```
=====================================================================
               Best Base Model Cluster 1 Summary
                Results: Ordinary least squares
=====================================================================
Model:              OLS              Adj. R-squared:     0.983
Dependent Variable: Withdraw         AIC:                478.1824
Date:               2024-11-09 19:50 BIC:                493.7710
No. Observations:   364              Log-Likelihood:     -235.09
Df Model:           3                F-statistic:        6884.
Df Residuals:       360              Prob (F-statistic): 1.87e-317
R-squared:          0.983            Scale:              0.21543
---------------------------------------------------------------------
            Coef.    Std.Err.     t      P>|t|    [0.025    0.975]
---------------------------------------------------------------------
Intercept  -28.3978    1.1959  -23.7458  0.0000  -30.7497  -26.0460
Shops        0.1324    0.0012  111.8871  0.0000    0.1301    0.1348
ATMs        -1.0044    0.0125  -80.3681  0.0000   -1.0290   -0.9798
High         1.0594    0.0531   19.9376  0.0000    0.9549    1.1639
```

```
---------------------------------------------------------------------
Omnibus:                4.066        Durbin-Watson:           2.069
Prob(Omnibus):          0.131        Jarque-Bera (JB):        3.946
Skew:                   0.255        Prob(JB):                0.139
Kurtosis:               3.033        Condition No.:           49121
=====================================================================

=====================================================================
                Best Enhanced Model Cluster 2 Summary
                    Results: Ordinary least squares
=====================================================================
Model:                OLS               Adj. R-squared:      0.977
Dependent Variable:   Withdraw          AIC:                 17457.8369
Date:                 2024-11-09 19:50  BIC:                 17605.6751
No. Observations:     11991             Log-Likelihood:      -8708.9
Df Model:             19                F-statistic:         2.639e+04
Df Residuals:         11971             Prob (F-statistic):  0.00
R-squared:            0.977             Scale:               0.25067
---------------------------------------------------------------------
            Coef.  Std.Err.    t    P>|t|   [0.025   0.975]
---------------------------------------------------------------------
Intercept                1.1096   0.0277 40.0325 0.0000   1.0553  1.1640
log_Shops                8.1379   0.5178 15.7155 0.0000   7.1229  9.1529
ATMs_log_Shops          -0.8183   1.7603 -0.4649 0.6420  -4.2688  2.6322
Workday                 -0.1566   0.1086 -1.4417 0.1494  -0.3696  0.0563
Center_log_Shops         0.3747   1.5610  0.2400 0.8103  -2.6851  3.4344
High_ShopsCubed          0.0000   0.0000  4.0697 0.0000   0.0000  0.0000
ShopsCubed_log_Shops     0.0000   0.0000  6.7273 0.0000   0.0000  0.0000
ATMs_Workday             0.0047   0.0052  0.9080 0.3639  -0.0054  0.0148
Center_log_ATMs         -0.8595   0.4921 -1.7467 0.0807  -1.8241  0.1050
ATMs_Center              0.0825   0.0519  1.5912 0.1116  -0.0191  0.1842
High_log_ATMs            0.2538   0.0680  3.7314 0.0002   0.1205  0.3872
High_ATMsCubed          -0.0001   0.0000 -4.3669 0.0000  -0.0002 -0.0001
ATMsCubed_log_ATMs      -0.0013   0.0005 -2.7793 0.0055  -0.0023 -0.0004
log_ATMs                -0.7031   0.2426 -2.8979 0.0038  -1.1787 -0.2275
ATMsSquared_ShopsCubed  -0.0000   0.0000 -4.9952 0.0000  -0.0000 -0.0000
Center_ShopsCubed       -0.0000   0.0000 -0.1654 0.8686  -0.0000  0.0000
Center                   0.0618   0.2570  0.2404 0.8100  -0.4420  0.5656
Workday_Center           0.0618   0.2570  0.2404 0.8100  -0.4420  0.5656
Workday_ShopsCubed      -0.0000   0.0000 -0.6177 0.5368  -0.0000  0.0000
Workday_log_Shops       -0.2567   0.0315 -8.1397 0.0000  -0.3185 -0.1949
ShopsSquared_ATMsCubed   0.0000   0.0000  3.4387 0.0006   0.0000  0.0000
log_Shops_log_ATMs      -4.0292   1.9222 -2.0962 0.0361  -7.7969 -0.2614
High_Workday             0.0038   0.0226  0.1659 0.8683  -0.0406  0.0481
Center_ShopsSquared      0.0000   0.0000  0.1523 0.8789  -0.0001  0.0001
Shops_log_ATMs           0.0256   0.0136  1.8832 0.0597  -0.0010  0.0522
ATMs                     5.8911  12.1151  0.4863 0.6268 -17.8564 29.6387
---------------------------------------------------------------------
Omnibus:              1.599      Durbin-Watson:         2.020
Prob(Omnibus):        0.449      Jarque-Bera (JB):      1.601
Skew:                 0.011      Prob(JB):              0.449
Kurtosis:             2.948      Condition No.:         15307874103660700
=====================================================================

=====================================================================
                Best Enhanced Model Cluster 3 Summary
                  Results: Ordinary least squares
=====================================================================
Model:                OLS               Adj. R-squared:      0.938
Dependent Variable:   Withdraw          AIC:                 7690.9742
```

```
Date:                2024-11-09 19:50 BIC:                7809.1447
No. Observations:    5245             Log-Likelihood:     -3827.5
Df Model:            17               F-statistic:        4632.
Df Residuals:        5227             Prob (F-statistic): 0.00
R-squared:           0.938            Scale:              0.25285
-------------------------------------------------------------------------
                    Coef.  Std.Err.    t     P>|t|   [0.025   0.975]
-------------------------------------------------------------------------
Intercept          10.1613   2.1907   4.6383 0.0000   5.8666 14.4561
ATMs_Workday        0.0115   0.0148   0.7815 0.4346  -0.0174  0.0405
Center             18.8237  17.6554   1.0662 0.2864 -15.7882 53.4356
ATMs               -0.8059   0.6191  -1.3018 0.1930  -2.0195  0.4077
High_ShopsSquared   0.0005   0.0004   1.1083 0.2678  -0.0004  0.0013
Workday             8.9761   7.7652   1.1559 0.2478  -6.2469 24.1991
Workday_Center     -1.9630   0.0483 -40.6832 0.0000  -2.0576 -1.8684
log_Shops           2.2564   0.4761   4.7397 0.0000   1.3231  3.1897
High_ShopsCubed    -0.0000   0.0000  -1.1011 0.2709  -0.0000  0.0000
High_log_ATMs       0.0057   0.0145   0.3922 0.6950  -0.0228  0.0341
High_Center        -0.0264   0.0490  -0.5394 0.5897  -0.1225  0.0696
Workday_log_Shops  -2.5580   1.8143  -1.4099 0.1586  -6.1147  0.9987
Workday_ShopsCubed  0.0000   0.0000   1.2432 0.2139  -0.0000  0.0000
High               -0.5689   1.4195  -0.4008 0.6886  -3.3517  2.2138
Center_ShopsCubed   0.0000   0.0000   0.8213 0.4115  -0.0000  0.0000
Center_log_Shops   -3.2607   4.1306  -0.7894 0.4299 -11.3584  4.8371
ATMs_log_Shops     -0.0442   0.1345  -0.3283 0.7427  -0.3079  0.2195
ATMs_Center         0.0026   0.0226   0.1133 0.9098  -0.0418  0.0469
-------------------------------------------------------------------------
Omnibus:              2.409       Durbin-Watson:          1.997
Prob(Omnibus):        0.300       Jarque-Bera (JB):       2.395
Skew:                 0.032       Prob(JB):               0.302
Kurtosis:             2.917       Condition No.:          2402506187
=========================================================================
```

Model	**CV Test MSE**
Best Base Model Cluster 1	0.2138
Best Base Model Cluster 2	0.2487
Best Base Model Cluster 3	0.3279
Best Enhanced Model Cluster 1	0.2189
Best Enhanced Model Cluster 2	0.2485
Best Enhanced Model Cluster 3	0.2524

Table 3: Cluster Models CV Test MSE Comparison

Discussion

Model Comparison

Model	Test MSE	R2
Polynomial	0.2477	0.9996
Ridge	0.2487	0.9996
Lasso (with k-means clustering)	0.2429 avg	0.9383 0.9745 0.9978 0.9758
Elastic net	0.2810	0.9996
Individual Cluster OLS	0.2382 avg	0.9830 0.9770 0.9380

Table 4: Model Comparison Table

Final Model

$$
\text{Withdraw} \sim \begin{cases} \text{Shops} + \text{ATMs} + \text{High} & \text{if Cluster 1} \\ \log(\text{Shops}) + \text{ATMs} \cdot \log(\text{Shops}) + \text{Workday} + \text{Center} \cdot \log(\text{Shops}) + \text{Shops} \cdot \text{High} + \text{Shops} \\ \quad + \text{Center} \cdot \log(\text{ATMs}) + \text{Workday} \cdot \log(\text{ATMs}) + \text{ATMs} \cdot \text{Center} + \text{Workday} \cdot \text{ATMs}^2 \\ \quad + \text{Center} + \text{Shops}^2 \cdot \text{ATMs}^3 + \text{ATMs} \cdot \text{Shops}^3 + \text{ATMs} + \text{ATMs}^3 \cdot \log(\text{ATMs}) \\ \quad + \text{High} \cdot \log(\text{Shops}) + \text{Workday} \cdot \text{Center} + \text{Center} \cdot \text{ATMs}^2 & \text{if Cluster 2} \\ \text{ATMs} \cdot \text{Workday} + \text{Center} + \text{ATMs} + \text{High} \cdot \text{Shops}^2 + \text{Workday} + \text{Workday} \cdot \text{Center} \\ \quad + \log(\text{Shops}) + \text{High} \cdot \text{Shops}^3 + \text{High} \cdot \log(\text{ATMs}) + \text{High} \cdot \text{Center} \\ \quad + \text{Workday} \cdot \log(\text{Shops}) + \text{Workday} \cdot \text{Shops}^3 + \text{High} + \text{Center} \cdot \text{Shops}^3 \\ \quad + \text{Center} \cdot \log(\text{Shops}) + \text{ATMs} \cdot \log(\text{Shops}) + \text{ATMs} \cdot \text{Center} & \text{if Cluster 3} \end{cases}
$$

where:

- Cluster 1: Downtown AND Holiday AND Center
- Cluster 2: Downtown AND Holiday AND Not Center OR Downtown AND Workday
- Cluster 3: Not Downtown

Model Interpretability
The piecewise regression model developed for predicting ATM withdrawals based on environmental and temporal factors provides a comprehensive view of how different features influence cash usage in distinct contexts. The model is divided into three clusters, each representing a specific set of conditions that capture the heterogeneity of ATM withdrawal patterns across different locations and times.

In Cluster 1, which corresponds to ATMs located in downtown areas, open on holidays, and situated in commercial centers, the primary drivers of withdrawal behavior are the number of shops, the presence of other ATMs, and the historical cash demand of the ATM. The positive relationships between shops and ATMs with withdrawals suggest that commercial areas with more shops and surrounding ATMs tend to see higher levels of cash withdrawn. This makes sense given that these areas often attract more visitors, especially during holidays, which leads to greater demand for cash. The inclusion of the High variable, indicating high past cash demand, further strengthens the interpretation that ATMs with a history of significant withdrawals are likely to continue to see similar demand, possibly due to sustained foot traffic or high consumer activity in these areas.

Cluster 2, which includes ATMs in downtown areas during either holidays or workdays and those located outside centers, displays a more complex relationship with its features. The logarithmic transformation of Shops and its interaction with ATMs suggests diminishing returns as the number of shops increases. This implies that, while

more shops generally correlate with higher withdrawal rates, the effect of adding additional shops decreases once a critical threshold is crossed. Similarly, the Workday variable introduces a different dynamic, showing that ATM usage patterns are affected by whether the day is a workday or a holiday, particularly in non-center locations. The Center feature further contributes to understanding withdrawal behavior in areas without central commercial hubs, indicating that, even in downtown locations, ATMs in non-center areas exhibit different withdrawal characteristics. The multiplicative terms, such as the interaction between Shops and High, suggest that areas with high commercial activity and a history of high cash demand are particularly important in driving withdrawal behavior, especially in downtown areas during holidays.

Finally, Cluster 3, which represents ATMs located outside downtown areas, is influenced more directly by the presence of ATMs and the Workday variable, with High still playing a significant role. The simplicity of this cluster reflects the fact that withdrawal behavior in non-downtown areas is less influenced by the commercial environment and more by the accessibility and past performance of the ATMs. The positive effect of ATMs suggests that when more ATMs are available in the vicinity, withdrawals tend to increase, possibly due to greater convenience for users. The importance of High indicates that historical patterns of cash demand remain a significant predictor, even in less commercialized areas. This underscores the relevance of past demand in predicting future withdrawal behavior, especially in locations where commercial factors like the number of shops are less influential.

Overall, the piecewise model provides valuable insights into the different dynamics that govern ATM usage. The key features—such as the number of Shops, the presence of ATMs, the Workday status, and the Center designation—interact in complex ways to influence withdrawal behavior. By segmenting the data based on location and temporal factors, the model accounts for the varied contexts in which ATMs operate, allowing for more accurate predictions and a deeper understanding of how different factors come together to affect cash withdrawals. This segmentation also highlights the importance of considering local conditions, such as the commercial activity in downtown areas or the workday status, to predict ATM usage more effectively.

Cautionary Notes
We would like to highlight several cautionary notes regarding the clustering approach and the overall modeling process.

Firstly, Cluster 1 is the best-performing cluster based on predictive accuracy, yet it represents a significantly small dataset compared to Cluster 2 and Cluster 3, with only 455 data points. This small sample size in Cluster 1 may lead to overfitting or a model that does not generalize well to new data. To enhance predictive performance in Cluster 1, alternative clustering methods such as hierarchical clustering could be explored. This method does not require a pre-defined number of clusters and may reveal additional sub-clusters based on finer details like the time of day or specific location (Hastie, Tibshirani, Friedman, 2009, p. 250). Hierarchical clustering could help uncover nuanced differences in holiday withdrawal patterns between business downtown areas and shopping districts, which are currently grouped together in Cluster 1.

Additionally, while the model's predictive performance is of paramount importance, we must acknowledge the complexity introduced by the high AIC and BIC scores. These metrics reflect the trade-off between model fit and complexity, with high values suggesting that the model might be too complex. Although this complexity may improve predictive accuracy, it comes at the expense of interpretability. The high number of features, interactions, and polynomial terms included in the model make it difficult to interpret and understand the individual contributions of each feature. This is a known trade-off when focusing on prediction, but it limits the model's usability for real-world decision-making where interpretability is critical.

Another limitation arises from the sensitivity to initial conditions in the clustering process. The results depend heavily on the initial number of clusters and their size. Small variations in these initial conditions could lead to different clusters, potentially affecting the model's stability and generalizability. Moreover, the model's performance might also be impacted by outliers in the data that were not removed. These outliers can distort the cluster assignment, leading to inaccurate predictions or a failure to capture the true patterns within the data. Further data cleaning or robust clustering techniques could mitigate these issues.

Finally, while the model incorporates valuable features such as the number of shops and ATMs, the role of local economic conditions and other external factors (e.g., weather, local events) have not been considered in this analysis. These factors could also influence ATM withdrawals, and their omission may limit the model's generalizability across different contexts. Further refinement of the model could involve exploring additional

variables to capture these external influences.

In summary, while the current model provides valuable insights into ATM withdrawal behavior, its complexity, dependence on clustering choices, and potential sensitivity to outliers and missing external factors require caution when interpreting the results and applying them in practice.

Asymmetric Loss
As mentioned in the introduction of this paper, there are two scenarios in which the bank can make a loss, either having excess funds in any given ATM, or alternatively, insufficient funds. This report has been centered on the minimisation of the MSE loss, but ultimately one of these scenarios is likely to result in higher financial loss to the bank. In the situation of excess funds, there exists an opportunity cost, i.e. some of the funds could be used for profit-generating activities rather than sitting idle. In the case of a lack of funds, as mentioned previously, there are operational costs to consider with replenishment and potential loss in customer satisfaction. This motivates the use of an asymmetric loss function, where prediction errors in estimating cash requirements for an ATM result in different levels of financial loss. Underestimating demand, for example, may incur a higher loss compared to overestimating, and hence greater penalties should be placed on having insufficient funds.

Appendix

Generalised Additive Models

Generalised Additive Models (GAMs) are nonparametric regression techniques. GAMs allow for more flexibility in functional form when compared with linear models and regularisation techniques, providing a visual interpretation of each predictor and its respective function estimate. GAMs have the following form:

$$E(Y|X_1, X_2, ..., X_p) = \alpha + f_1(X_1) + f_2(X_2) + ... + f_p(X_p)$$

Each function is an unspecified function, generally fitted using splines, but linear and factor terms can be included as well (Hastie, Tibshirani and Friedman, 2009, p.219, 295-298). GAMs are not comparable to the other models mentioned in this project in terms of test MSE, but they offer perhaps a better perspective on how each predictor affects its corresponding fitted function.

The python package pyGAM was used for this modelling, specifically LinearGAM, which specifies the errors in the model as normally distributed, along with an identity link function. Only the base covariates were included in the model, not including Downtown, which was removed due to multicollinearity issues. Cross validation with five folds was used to measure test MSE, and the model specified used splines for Shops and ATMs, then factor terms for Workday, Center, and High. Interaction terms were not added due to this modelling being an extension to the project, better utilised for aid in interpretation of the relationships between cash demand and the covariates, rather than prediction. After specifying the functional type of each predictor, within each fold, gridsearch was employed to find the optimal smoothing parameter lambda, subsequently the test MSE is found in each fold. Average test MSE over the folds ended up being 6.2041, significantly larger than all other models discussed, but GAMs do give the following output:

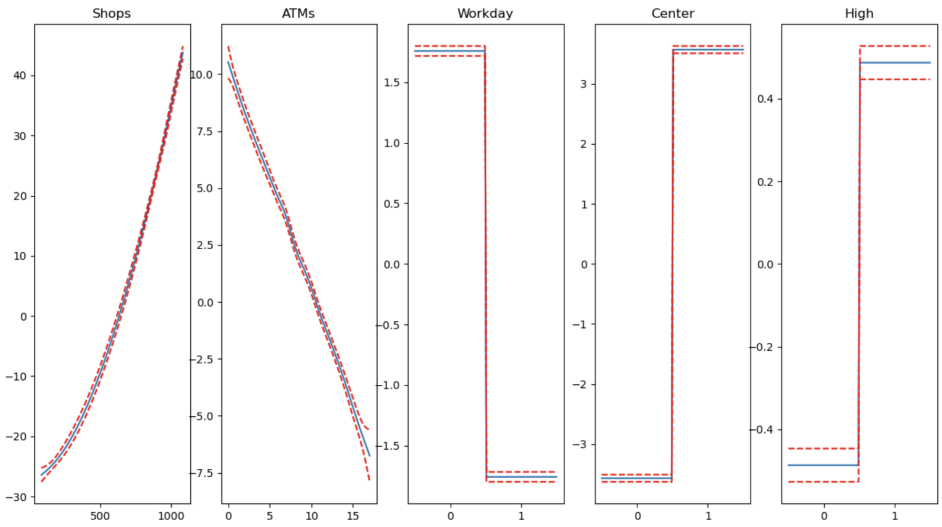


Figure 8

These plots represent the partial effects of each predictor on The confidence intervals are fairly narrow for most estimates apart from High, and some slight nonlinearity is clear in both Shops and ATMs. For Shops the confidence intervals widen at the base of the function, suggesting certainty in the estimate, likely due to the concentration of points in that region and the distribution. It's clear that as the number of shops increases, the value of the function increases and hence cash withdrawals increase, this aligns with all the previous modelling, but gives more insight when compared to the single point estimates produced by the previous models. For ATMs, it's evident that more ATMs in walking distance leads to greater dispersion in withdrawals, so for any one ATM, cash demand is decreased. The factor variables are fairly intuitive, e.g. if the day is a holiday, cash demand is increased in comparison to work days.

To conclude, whilst GAMs are a great tool for visualisation, they don't aid in prediction enough to warrant utilising as one of the primary models, as this is primarily a prediction problem, which is why they have been included in the appendix.

References

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed., Chapter 7: Model Assessment and Selection, p. 219, 250, Chapter 9: Additive Models, Trees, and Related Methods, 295–298, Chapter 14: Unsupervised Learning, p. 510.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in Python.* 3rd ed., Chapter 3: Linear Regression, p. 108, Chapter 6: Linear Model Selection and Regularization, 240 - 244.

Scikit-learn Developers. *Scikit-learn: Machine Learning in Python.* Available at: https://scikit-learn.org

Scikit-learn. *ElasticNetCV.*
Available at: https://scikit-learn.org/dev/modules/generated/sklearn.linear_model.ElasticNetCV.html