Module 5 Final Project Draft Report — Bank Marketing

Muhammad Umer Mirza

College of Professional Studies, Northeastern University Toronto

ALY6015 - Intermediate Analytics

Dr. Matthew Goodwin

February 11, 2024

Introduction

In this report, we dive deeper into the bank marketing dataset using advanced statistical techniques. We explore the potential of regularization methods, such as Ridge and Lasso regression, and Stepwise regression, to optimize our analysis and identify the most significant predictors of term deposit subscriptions. We also introduce a non-parametric test to broaden our statistical approach and unearth more intricate patterns within the data. By employing these sophisticated methods, we aim to improve our strategy for bank marketing decisions, with more refined insights and informed decision-making.

Exploratory Data Analysis

Descriptive Statistics for Numerical Variables

Variable	Min	Max	Mean	SD	Median	Q1.25%	Q3.75%
age	19	87	41.17	10.58	39	33	49
balance	-3313	71188	1422.66	3009.64	444	69	1480
contact_date	1	31	15.92	8.25	16	9	21
duration	4	3025	263.96	259.86	185	104	329
campaign	1	50	2.79	3.11	2	1	3
pdays	-1	871	39.77	100.12	-1	-1	-1
previous	0	25	0.54	1.69	0	0	0

Figure 1: Descriptive Statistics for Numerical Variables.

Figure 1 provides a comprehensive overview of the data set, highlighting the varying characteristics across client demographics and account characteristics. The age of clients ranges from 19 to 87 years, indicating a wide range of ages that may influence their decision to subscribe to term deposits. The account balance variable exhibits a large standard deviation of 3,009.64, indicating a diverse range of financial backgrounds. This information can be instrumental in predicting term deposit interest. The contact duration variable ranges from a minimum of 4 to a maximum of 3,025 seconds, suggesting varying levels of client engagement. Notably, the 'pdays' feature shows that 75% of observations are at -1, indicating that many clients have yet to be contacted before, highlighting a potentially untapped market. These detailed numerical insights are critical in understanding customer profiles, which can help fit a model predicting the likelihood of subscribing to a term deposit based on demographic and account characteristics.

Summary for Specified Categorical Variables

Variable	Level	Count	Percentage.Freq
marital_status	married	2797	61.87
marital_status	single	1196	26.45
marital_status	divorced	528	11.68
education	secondary	2306	51.01
education	tertiary	1350	29.86
education	primary	678	15.00
education	unknown	187	4.14
credit_default	no	4445	98.32
credit_default	yes	76	1.68
housing_loan	yes	2559	56.60
housing_loan	no	1962	43.40
loan	no	3830	84.72
loan	yes	691	15.28
contact_type	cellular	2896	64.06
contact_type	unknown	1324	29.29
contact_type	telephone	301	6.66
poutcome	unknown	3705	81.95
poutcome	failure	490	10.84
poutcome	other	197	4.36
poutcome	success	129	2.85
subscribe_term_deposit	no	4000	88.48
subscribe_term_deposit	yes	521	11.52

Figure 2: Descriptive Statistics for Specified Categorical Variables.

Figure 2 presents the distribution of key categorical variables that inform our predictive model. Marital status shows that the majority of clients are married (61.87%), followed by single (26.45%) and divorced (11.68%) individuals. In terms of education, most clients have completed secondary education (51.01%). Notably, a vast majority have no credit default (98.32%), and more than half possess a housing loan (56.60%). The preferred contact type is overwhelmingly cellular (64.06%). The outcome of the previous marketing campaign was unknown for most (81.95%), with a small fraction marking success (2.85%). These categorical trends, alongside the numerical data, provide a comprehensive view for modeling term deposit subscription likelihood, with particular attention to the influence of marital status, education, and financial commitments.

Analysis

Our analysis seeks to answer the question: Can the likelihood of a customer subscribing to a term deposit be predicted from their demographic profile and account characteristics? We are examining variables such as age, job title, marital status, education, credit default status, account balance, and existing loans. To approach this, we employ regularization techniques using Ridge and Lasso regression via the cv.glmnet function and compare them with stepwise model selection. The objective is to use these regularization methods to discern the relationships among the variables and to make accurate predictions regarding term deposit subscriptions.

Why Regularization Methods and Stepwise Model Selection

Incorporating regularization methods like Ridge and Lasso regression alongside stepwise model selection offers a comprehensive approach to predicting customer subscription to term deposits. Regularization methods effectively address multicollinearity and overfitting by penalizing large coefficients and enabling feature selection, leading to more stable and generalizable models. Stepwise model selection complements this by iteratively adding or removing predictors based on statistical criteria, helping to refine the model further. This combination allows for a nuanced comparison of methodologies, leveraging the strengths of both regularization for managing complex data relationships and stepwise selection for its straightforward criteria-based model simplification. These techniques are suitable for analyzing the intricate relationships within our dataset and enhancing the reliability of our predictions.

Split Data Into Train and Test Sets

To establish a robust foundation for our predictive models, we initiated our analysis by creating distinct training and testing datasets. This critical step ensures that our models are evaluated on fresh data, simulating real-world performance. Our process for the split adhered to the guidelines outlined in the Feature Selection Document, where 70% of the data was allocated for training, and the remaining 30% for testing. The feature matrix for each set was constructed to include the chosen variables—age, job title, marital status, education, credit default, balance, housing loan, and personal loan—which are pivotal in forecasting term deposit subscriptions. This structured division of data facilitates an unbiased assessment of the predictive power of our models.

Ridge Regression

```
> log(cv_fit_ridge$lambda.min) # Optimal for prediction
[1] -1.618258
> log(cv_fit_ridge$lambda.1se) # Within one standard error
[1] -0.2227523
```

Figure 3: Ridge Regression Lambda min and 1.se values.

In Ridge regression, the lambda parameter controls the amount of shrinkage: the larger the lambda, the more shrinkage occurs. The optimal lambda for prediction, lambda.min, is calculated to be exp(-1.618258), indicating the value at which the model will likely perform best at making predictions. On the other hand, lambda.1se is exp(-0.2227523), representing a more regularized model within one standard error of the minimum. It trades off some of the model's predictive power for greater robustness and simplicity. Comparing these values, lambda.min offers the least bias, while lambda.1se provides a more conservative model that may perform better when generalizing to new data.

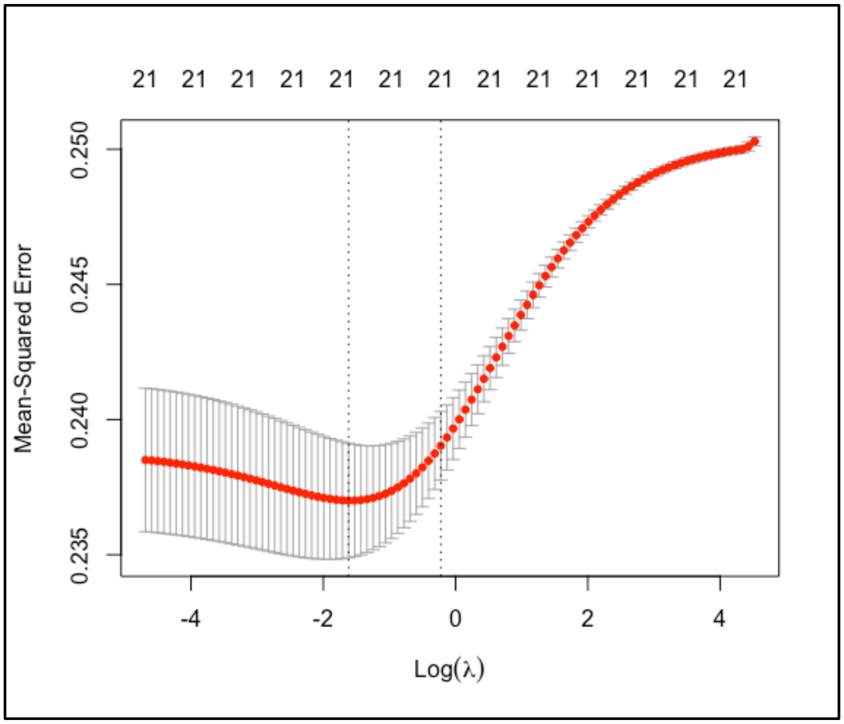

Figure 4: Ridge Regression Plot.

Figure 4 displays a cross-validation plot for selecting the lambda parameter in Ridge regression. The x-axis represents the log-transformed lambda values, while the y-axis shows the mean squared error (MSE) for each lambda. The dotted vertical lines mark the lambda.min and lambda.1se values. The red dots indicate the MSE for each model, with the red line connecting them, showing the error trend. The optimal point, lambda.min, is where the MSE is at its lowest, suggesting the best model fit. The lambda.1se provides a model with potentially greater generalizability by introducing more bias in exchange for reduced variance.

```
> coef(model_ridge_min)
22 x 1 sparse Matrix of class "dgCMatrix"
                                     s0
(Intercept)             0.1350203163874
age                     0.0004427605502
job_titleblue-collar   -0.0238954222311
job_titleentrepreneur  -0.0093075183991
job_titlehousemaid      0.0037998049308
job_titlemanagement     0.0072322677704
job_titleretired        0.0681621678002
job_titleself-employed  0.0038733581452
job_titleservices      -0.0122017492970
job_titlestudent        0.0482379216146
job_titletechnician    -0.0063981887855
job_titleunemployed    -0.0186618056155
job_titleunknown       -0.0038200499173
marital_statusmarried  -0.0212305370608
marital_statussingle    0.0098146597198
educationsecondary      0.0000203850803
educationtertiary       0.0161148782631
educationunknown       -0.0318732877719
credit_defaultyes       0.0033049455823
balance                -0.0000003755506
housing_loanyes        -0.0388069760944
loanyes                -0.0372337437143
```

```
> coef(model_ridge_1se)
22 x 1 sparse Matrix of class "dgCMatrix"
                                      s0
(Intercept)             0.12337806282457
age                     0.00026632333892
job_titleblue-collar   -0.01367933100744
job_titleentrepreneur  -0.00507788661036
job_titlehousemaid      0.00315138706178
job_titlemanagement     0.00576507803680
job_titleretired        0.03443543298283
job_titleself-employed  0.00261070682537
job_titleservices      -0.00666472774665
job_titlestudent        0.02573036417465
job_titletechnician    -0.00229196734838
job_titleunemployed    -0.00731037764593
job_titleunknown       -0.00007702572334
marital_statusmarried  -0.01073602300366
marital_statussingle    0.00730464124273
educationsecondary     -0.00249879179632
educationtertiary       0.00916682426081
educationunknown       -0.01349393860191
credit_defaultyes       0.00105436566263
balance                 0.00000001306185
housing_loanyes        -0.01961117084737
loanyes                -0.01742230280039
```

Figure 5: Ridge regression lambda.min and lambda.1se model coefficients.

The coefficients from the Ridge regression models provide insights into the factors influencing the likelihood of a customer subscribing to a term deposit. For the model_ridge_min, the most notable positive coefficient is for job_titleretired, suggesting retirees may be more inclined to subscribe. Interestingly, job_titlestudent also shows a positive association, indicating students as potential subscribers. The negative coefficients for housing_loanyes and loanyes suggest that having loans is associated with a lower likelihood of subscription. The model_ridge_1se coefficients are generally smaller in magnitude, reflecting a more conservative model, but the trends remain similar. The positive influence of being retired or a student and the negative impact of loans on subscription likelihood are consistent findings in both models.

```
> cat("RMSE Training Ridge:", rmse_train_ridge, "\n")
RMSE Training Ridge: 0.3173201
> cat("RMSE Test Ridge:", rmse_test_ridge, "\n")
RMSE Test Ridge: 0.3141094
```

Figure 6: Ridge Regression RMSE values for train and test sets.

The Root Mean Square Error (RMSE) for both the training and test datasets in the Ridge regression model are quite close, with the training set at 0.3173201 and the test set at 0.3141094. This proximity in RMSE values suggests that the model is generalizing well to new data, indicating that there is no significant overfitting. Overfitting would typically be characterized by a low RMSE on the training set and a much higher RMSE on the test set. Here, the consistency between the two indicates a stable model performance.

LASSO

```
> log(cv_fit_lasso$lambda.min) # Optimal for prediction
[1] -4.71163
> log(cv_fit_lasso$lambda.1se) # Within one standard error
[1] -3.595225
```

Figure 7: LASSO Lambda min and 1.se values.

The Lasso regression model's optimal lambda value for prediction, lambda.min, is given by exp(-4.71163), which is the value that minimizes the cross-validated mean squared error. The lambda.1se value, exp(-3.595225), is the more regularized model that is within one standard error of the minimum. A larger absolute value of log(lambda) for lambda.min compared to lambda.1se suggests a greater level of shrinkage on the coefficients, leading to a sparser model. This can often result in a model that retains only the most significant predictors, potentially enhancing interpretability and reducing the risk of overfitting.

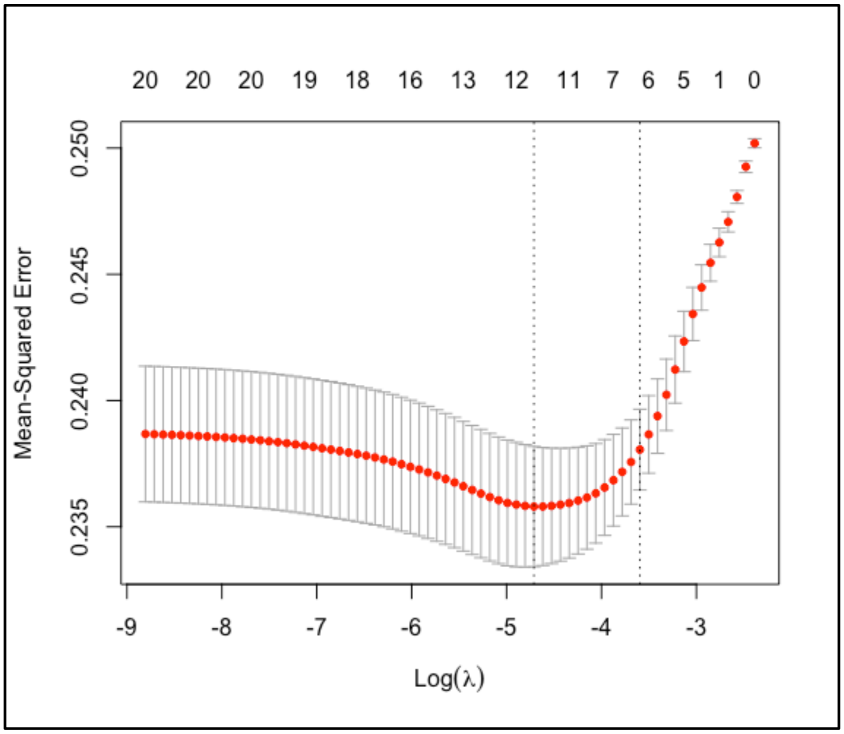

Figure 8: LASSO plot.

Figure 8 showcases the Lasso regression's cross-validation results. As we adjust the regularization strength, indicated by log-transformed lambda values, we observe the model's mean squared error (MSE) reacting correspondingly. The plot reveals a minimum MSE at the lambda.min point, where the model retains 12 variables, suggesting a detailed representation of the data. As we move towards the lambda.1se point, indicating a more regularized model, the number of variables retained drops to 6, highlighting the most substantial predictors for a more parsimonious model. This strategic reduction could enhance the model's generalizability without a significant increase in error.

```
> coef(model_lasso_min)
22 x 1 sparse Matrix of class "dgCMatrix"
                                 s0
(Intercept)             0.157873653
age                     .
job_titleblue-collar   -0.014912035
job_titleentrepreneur   .
job_titlehousemaid      .
job_titlemanagement     .
job_titleretired        0.074751727
job_titleself-employed  .
job_titleservices       .
job_titlestudent        0.002788892
job_titletechnician     .
job_titleunemployed     .
job_titleunknown        .
marital_statusmarried  -0.022532619
marital_statussingle    .
educationsecondary      .
educationtertiary       0.015145367
educationunknown       -0.002531061
credit_defaultyes       .
balance                 .
housing_loanyes        -0.048540691
loanyes                -0.035902032
```

```
> coef(model_lasso_1se)
22 x 1 sparse Matrix of class "dgCMatrix"
                                s0
(Intercept)             0.12849179
age                     .
job_titleblue-collar    .
job_titleentrepreneur   .
job_titlehousemaid      .
job_titlemanagement     .
job_titleretired        .
job_titleself-employed  .
job_titleservices       .
job_titlestudent        .
job_titletechnician     .
job_titleunemployed     .
job_titleunknown        .
marital_statusmarried   .
marital_statussingle    .
educationsecondary      .
educationtertiary       .
educationunknown        .
credit_defaultyes       .
balance                 .
housing_loanyes        -0.02157136
loanyes                 .
```

Figure 9: LASSO regression lambda.min and lambda.1se model coefficients.

In the Lasso regression outputs, several coefficients are reduced to zero, which indicates that these variables are not contributing to the model. In the model_lasso_min output, only age, job titles other than 'blue-collar', 'retired', and 'student', marital statuses other than 'married', and education levels other than 'tertiary' and 'unknown' are reduced to zero. In the model_lasso_1se output, a more conservative model, almost all coefficients have been shrunk to zero except for the intercept and the coefficient for housing_loanyes. This suggests that, at this level of regularization, only the variable associated with having a housing loan is considered a significant predictor in the context of the data and the specific LASSO model.

```
> cat("RMSE Training Lasso:", rmse_training_lasso, "\n")
RMSE Training Lasso: 0.3194983
> cat("RMSE Testing Lasso:", rmse_testing_lasso, "\n")
RMSE Testing Lasso: 0.3157283
```

Figure 10: LASSO Regression RMSE values for train and test sets.

The RMSE values for the Lasso regression model are quite similar for both the training set (0.3194983) and the test set (0.3157283). This small difference between the training and testing error indicates that the model is generalizing well to unseen data. Such a result suggests that there is no significant overfitting occurring with this model; overfitting would be indicated by a much lower RMSE on the training set compared to the test set. Therefore, the model's performance is stable across both datasets.

Stepwise Model Selection

```
> stepwise_model <- step(lm(subscribed ~ ., data = training_set), direction = 'both')
Start:  AIC=-8301.22
subscribed ~ age + job_title + marital_status + education + credit_default +
    balance + housing_loan + loan + contact_type + contact_date +
    contact_month + duration + campaign + pdays + previous +
    poutcome

                 Df Sum of Sq    RSS     AIC
- age             1     0.009 223.36 -8303.1
- previous        1     0.016 223.37 -8303.0
- campaign        1     0.025 223.38 -8302.9
- pdays           1     0.049 223.40 -8302.5
- balance         1     0.083 223.44 -8302.0
- education       3     0.385 223.74 -8301.8
- credit_default  1     0.122 223.48 -8301.5
<none>                        223.35 -8301.2
- job_title      11     1.712 225.07 -8299.1
- marital_status  2     0.450 223.81 -8298.8
- housing_loan    1     0.312 223.67 -8298.8
- contact_date    1     0.424 223.78 -8297.2
- loan            1     0.522 223.88 -8295.8
- contact_type    2     1.256 224.61 -8287.5
- contact_month  11    11.525 234.88 -8164.0
- poutcome        3    12.979 236.33 -8128.5
- duration        1    52.815 276.17 -7631.6
```

Figure 11: Final Stepwise Selection Model.

The stepwise selection method has refined our model to what is considered optimal based on the Akaike Information Criterion (AIC). The final model includes variables such as job title, marital status, housing loan, loan, contact type, contact date, contact month, duration, and poutcome, which collectively provide the best balance between the number of predictors and the model's ability to make accurate predictions. Variables that had a less significant impact on the likelihood of a customer subscribing to a term deposit have been removed. This streamlined model is now ready to be used for prediction.

```
> rmse_training_stepwise
[1] 0.2660988
> rmse_testing_stepwise
[1] 0.278817
```

Figure 12: RMSE value for Stepwise selection model.

The RMSE values for the stepwise regression model indicate good performance, with the training set achieving an RMSE of 0.2660988 and the testing set slightly higher at 0.278817. The proximity of these values suggests that the model is consistent and not overfitting. The model seems to generalize well, maintaining its predictive accuracy on unseen data.

Conclusion

Ridge regression, with RMSEs of 0.3173 (training) and 0.3141 (test), demonstrates a robust predictive capability, striking an excellent balance between model complexity and generalizability. It particularly shines by identifying key demographics like retirees and students as more likely to subscribe to term deposits, offering valuable insights directly relevant to our core question.

Lasso regression, though slightly less performant with RMSEs of 0.3195 (training) and 0.3157 (test), excels in model simplification by reducing less impactful variables to zero. This method is advantageous for interpretability and focusing on the most significant predictors.

The stepwise model, despite its lower training RMSE of 0.2661, shows a slightly higher test RMSE of 0.2788. This indicates a good fit but suggests a potential for overfitting compared to Ridge when considering the difference in RMSE values.

Given our goal to predict term deposit subscriptions based on specific characteristics, Ridge regression appears to offer the best model. It not only performs well in both training and test scenarios but also provides actionable insights into which customer segments are more likely to engage, making it a particularly valuable tool for addressing our fundamental question.

References

1. Bluman, A. (2018). Elementary Statistics: A Step by Step Approach (10th ed.). McGraw Hill. ISBN 13: 978-1-259-755330.
2. Kabacoff, R. (2011). R in Action. Manning (2nd ed.). Manning Publications Co. ISBN 978-1-935-182399.
3. Goodwin, M. (2024). Module 4. Canvas. https://northeastern.instructure.com/courses/164840/modules
4. Goodwin, M. (2024). Module 4 Pre-Assingment Lab. Canvas. https://northeastern.instructure.com/courses/164840/assignments/2098519
5. Bhattacharyya, S. (2018). Ridge and Lasso Regression: A Complete Guide with Python Scikit-Learn. Medium; Towards Data Science. https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b
6. Shah, R. (2021). Comparision of Regularized and Unregularized Models. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/08/performance-comparision-of-regularized-and-unregularized-regression-models/