Modelling and Predicting Mortgage Customer Retention with Binary Logistic Regression

Alyssa Pennini
169029459
Further Statistics MA7406
Professor Wang

1. SUMMARY

A bank, looking to prevent the loss of mortgage customers at the end of the 'lock-in' period, asked for an analysis of customers' personal circumstances to determine if any variables significantly influence the customer's decision to stay or leave the bank. An analysis was performed using binary logistic regression to model the bank data and model selection was performed using forward stepwise selection with assessment criteria of Akaike Information Criterion, Bayesian Information Criterion, McFadden R-squared, and Classification Tables. The analysis resulted in a model that shows a significant relationship between a customer's account status and their choice in being a branch user, a direct/telephone user and having an agent.

2. INTRODUCTION

A bank is looking to understand the possible reasons why mortgage customers may choose to leave after the 'lock-in' period of their mortgage deal ends. They wish to understand what could be done differently in an effort to ultimately improve their customer retention in the future. The problem is multifaceted due to the range in customer personal profiles as well as the aspects pertaining to their specific mortgage deals.

Provided with the personal data and bank details of 399 of their customers, a request was made to analyse the given data to determine if any personal circumstances are significantly associated with a customer's decision to stay or leave the bank. Limited to the application of generalised linear models, various combinations of the twenty-two personal attributes are modelled using binary logistic regression to test their relationship with the account status response variable.

Forward stepwise selection is the systematic process used to choose a model with Akaike Information Criterion, Bayesian Information Criterion, McFadden R-Squared, and Classification Tables being the main criteria used to assess the model fit. Over forty models are created and tested against the assessment criteria, resulting in a best-fit model with an accuracy of 92.23%, consisting of three predictor variables including Branch User, Agent, and Direct User.

3. METHODS

3.1. Generalised Linear Models

The scope of this analysis was limited to the use of Generalised Linear Models by which a response variable is modelled using explanatory or predictor variables for which information is available. The structure of generalised linear models is comprised of a random component which is the response variable being modelled, a systematic component of linear predictors, which consists of the explanatory variables and their coefficients, and a link function which relates the explanatory variables with the response variable (Agresti, p.3).

There are different link functions based on the probability distributions of the available response data, so the first step in creating a model is to determine which distribution family best describes the given data.

As the goal of this analysis is to determine the predictors most influential in a customer's decision to retain their mortgage account, a binomial distribution was chosen as representative of the bank data given that the response variable has a binary outcome (Agresti, p.5). Therefore, the generalised linear model used in this analysis is binary logistic regression in the form

$$logit(u_i) = \log\left(\frac{u_i}{1-u_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

where $\beta_0, \beta_1, \ldots, \beta_k$, are the coefficients and $x_{i1}, x_{i2}, \ldots, x_{ik}$ are the explanatory or predictor variables that create the systematic component. The resulting value is the log-odds of a successful outcome where the log-odds increase if the coefficients are greater than 0 and decrease if the coefficients are less than 0. The probability of a successful outcome can then be calculated using the form

$$u_i = \frac{exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}.$$

For binary logistic regression, the response variable is binary while the independent explanatory variables can be discrete, continuous, or a combination of both (Penn State, Lesson 6.2).

3.2. Assessment of Model Fit

The selection of the best-fit model is a useless pursuit without a means to assess the accuracy and fit of a given model and how it compares to other models. As the explanatory variables for this analysis are both discrete and continuous, certain assessment tools do not apply in some cases, specifically goodness-of-fit test statistics and p-values in relation to continuous data. In addition, as this analysis has a wide range of variables under consideration, a means to compare models that are not nested and do not share common variables is of great importance.

Therefore, the four main criteria for assessing model fit here are Akaike Information Criterion, Bayesian Information Criterion, McFadden R-Squared and Classification Tables. In the event a model consists of only categorical variables, then residual deviance and p-values are assessed to determine fit.

3.2.1. Akaike Information Criterion

The main criterion assessing fit is Akaike Information Criterion (AIC) which "is an estimate of a constant plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model, so that a lower AIC means a model is

considered to be closer to the truth" (The Methodology Centre). Therefore, the goal is to minimise the AIC value. However, let it be noted that AIC is not indicative of fit overall, just in a comparison between models (Andale). But it is especially useful in this analysis because it does not depend on the models to be nested for a comparison to be valid.

3.2.2. Bayesian Information Criterion

Similar to Akaike Information Criterion, Bayesian Information Criterion (BIC) "is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup, so that a lower BIC means that a model is considered to be more likely to be the true model" (The Methodology Centre). It is a stricter criterion, attaching a higher penalty than AIC for more parameters. Therefore, both AIC and BIC should be in agreement and together provide a better assessment of model fit.

3.2.3. McFadden R-Squared

The coefficient of determination, R-Squared, is a measure of the predictive relationship between the explanatory predictors and the response variable. As R-Squared tends to increase when parameters are added to a model regardless of fit, the use of adjusted R-Squared is a better option. "The adjusted R-Squared increases only if the new term improves the model more than would be expected by chance" (Frost, 2013), therefore providing a better indication of fit. Given that R-Squared and adjusted R-Squared are not available for use with logistic regression, McFadden's pseudo-R-Squared provides similar results and is used in its place (Alice).

3.2.4. Classification Table

A classification table is a 2 x 2 table comparing the actual values provided in the data set and the values predicted using the model. The cells of the 2 x 2 table consist of the number of True Positive, True Negative, False Positive and False Negative predictions. The accuracy of the model can be determined by summing the True Positives and True Negatives and dividing by the sum of all values in the table (Analytics Vidhya).

3.2.5. Residual Deviance and P-Value

Goodness-of-fit test statistics such as G^2 and Pearson X^2 depend on the residual deviance and degrees of freedom in order to test the level of association between the explanatory and response variables using a chi-squared limiting distribution. Though this distribution "does not occur for ungrouped data" (Agresti, p.181) and is, therefore, inapplicable for continuous variables, it is useful in the comparison of nested models. "The difference of deviances is the likelihood-ratio test statistic for comparing the models. If the simpler model holds, this difference has an approximate chi-squared distribution with degrees of freedom equalling the difference in parameters between the complex and simpler models" (Agresti, p.181).

3.3. Model Selection

In addition to assessment criteria, the problem of overfitting the data is a strong concern when selecting a model. According to Agresti in *Foundations of Linear and Generalized Linear Models*, "The model should be complex enough to fit the data well. On the other hand, it should smooth rather than overfit the data and ideally be relatively simple to interpret" (Agresti, p.143). There is a simple rule to prevent overfitting and it refers to the number of parameters included in the selected model. When looking at the data, the general rule is to have ten to fifteen times as many observations with the outcome of interest as there are parameters in the model (Frost, 2015). For example, in this analysis, there are ninety-three cases with the outcome of interest which means a maximum of six to nine parameters may be included in the final model.

The process of choosing a model during this analysis is a modified forward stepwise selection. Traditionally, "Forward selection adds terms sequentially. At each stage it selects the term giving the greatest improvement in fit…The process stops when further additions do not improve the fit, according to statistical significance or a criterion for judging the model fit (such as the AIC)" (Agresti, p.144).

The null model provides the initial comparison for AIC and BIC values. Beginning with a single predictor, a set of models are created by adding each explanatory variable to the null model to determine whether or not their presence improves the fit of the model. For both AIC and BIC, models over the respective minimum are eliminated from further analysis. The minimum AIC and BIC for the set, as well as the maximum R-Squared and calculated accuracy percentages, are noted and used as the comparison values for the next set of models.

Two-predictor models are then created using all combinations of the remaining single predictor models. The same procedure is performed, only retaining the combinations of variables which meet at least three of the four criteria, eliminating all others and updating the new minimums and maximums for comparison in the next set. For three-predictor models and above, only combinations of the previous models are created when one or more common variables is present.

Once a model is found that crosses the threshold of the minimum AIC and BIC, as well as the maximum R-Squared and accuracy rate, all variations of the interaction terms are added and assessed by the same criteria to determine if any interactions between variables improve the fit. In addition, if the model only contains categorical variables, the difference in residual deviance and the associated p-values are reviewed for the nested models.

3.4. Prediction Ability

Using the caTools library in R, the data are randomly split into training and testing sets using a specified split ratio. Each model is then fitted using the training data before creating a

classification table to determine training set accuracy rates. The fitted model is then used to predict the results of the testing set to determine prediction accuracy rates. The procedure is repeated multiple times for both sets, each with a new random sample, before being averaged and used as the training accuracy and testing prediction rates. (Alice, Analytics Vidhya)

4. ANALYSIS

4.1. Data Cleaning and Preparation

The data provided by the bank include information for 399 of their customers with twenty-three attributes describing the personal circumstances of the customer. There are four attributes deemed redundant in this analysis: Scheme Details, Account Number, Start Date and End Date. The information described by Scheme Details is included in both Scheme Name and Scheme Code, so its inclusion would not provide any additional information. Account Number is an identifier and unique to each account and therefore irrelevant. Finally, the applicable information associated with the Start and End Dates is already included in the analysis within the Loan Term attribute. With the exclusion of those four attributes, eighteen potential explanatory variables are left, consisting of both discrete and continuous variables, the classification of which is shown in Table 1.

Let it be noted that the continuous variables could be grouped together, but for this analysis, those variables were left untouched.

Table 1 – Classification of Variables Under Consideration

Discrete Variables (Factors in R)	
Holder – 2 levels	Customer Gender – 2 levels
Post code – 195 levels	Branch User – 2 levels
Loan Type – 4 levels	Direct User – 2 levels
Scheme Code – 10 levels	Internet – 2 levels
Scheme Name – 10 levels	Agent – 2 levels
Region – 10 levels	Acron Code – 43 levels
Continuous Variables (Integer or Number in R)	
Original Loan Amount	Customer Age
Loan Balance	Customer Membership
Loan Term	Customer Salary

There are two instances in which missing values were of concern, namely within the Gender and Acron Code variables. Gender has four missing values which were filled with the median/mode of F for female. The Acron Code variable has nineteen missing values which were just labelled as ‘Blank’ and counted as another level associated with the variable (Alice).

4.2. Single Predictor Model

The first step in the analysis is to establish the null model for the data which includes no explanatory variables, just an intercept. The null deviance, degrees of freedom (DF), the Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC) are noted for comparison with the rest of the potential models as seen in Table 2.1.

Table 2.1 – Details for the Null Model to be Used for Comparison
(Extract of Table 2 in Appendix)

Model	Response	Explanatory Variable	Null Deviance	Null DF	AIC	BIC
0	Status	Null	433.29	398	435.29	439.28

Eighteen single predictor models are created, each adding one of the possible explanatory variables to the null model and are analysed to determine if the variables are significant. Based solely on the Akaike Information Criteria, there are four single predictor models with an AIC higher than the null model, as seen in Table 2.2, which means that they produce a worse fit and should not be included any further in the analysis.

Table 2.2 – Single Predictor Models with AIC above Null Model (435.29)
(Extract of Table 2 in Appendix)

Model	Response Variable	Explanatory Variable	Residual Deviance	Residual DF	AIC
1	Status	Holder	433.29	397	437.29
4	Status	Amount	432.96	397	436.96
12	Status	Gender	433.29	397	437.29
15	Status	Internet	431.69	397	435.69

Assessing the models according to the stricter Bayesian Information Criteria results in six additional models, shown in Table 2.3, being rejected for having a BIC higher than the null model.

Table 2.3 – Single Predictor Models with BIC above Null Model (439.28)
(Extract of Table 2 in Appendix)

Model	Response Variable	Explanatory Variables	Residual Deviance	Residual DF	BIC
2	Status	Post	17.32	204	1185.16
3	Status	Type	422.37	395	446.32
7	Status	Balance	431.07	397	443.04
10	Status	Region	402.12	389	462.01
11	Status	Membership	429.38	397	441.36
17	Status	Acron	315.57	356	573.09

Due to the concern regarding overfitting, the variables Scheme Name and Scheme Code are eliminated as both of their single predictor models contain ten parameters each which exceed the absolute maximum of nine parameters.

As a result of AIC, BIC and the rule to avoid overfitting, twelve models are eliminated from further consideration, leaving six possible explanatory variables – Term, Age, Branch, Direct, Agent and Salary – to be included in further analysis. Of these six variables, Agent provides the minimum AIC and BIC as well as the maximum accuracy and McFadden R-Squared as shown in the Tables 2.4 and 3.1 below. The full tables are available in the appendix.

Table 2.4 – Minimum AIC and BIC for Single Predictor Models
(Extract of Table 2 in Appendix)

Explanatory Variable	AIC	BIC
Agent	323.59	331.57

Table 3.1 – Maximum Accuracy and R-Squared for Single Predictor Models
(Extract of Table 3 in Appendix)

Explanatory Variable	R-Squared	Accuracy
Agent	0.2624093	0.7669173

4.3. Two-Predictor Model

Given Term, Age, Branch, Direct, Agent and Salary variables, fifteen combinations are possible in constructing two-predictor models. However, only three models fall below the minimum AIC and BIC thresholds, shown in Table 4.1, as well as being equal to or greater than the maximum accuracy and McFadden R-Squared values. Those results are available in Table 5 in the appendix. One additional model meets three out of four of the assessment criteria and is included in the table below.

Table 4.1 – Two-Predictor Models with Minimum AIC
(Extract of Table 4 in Appendix)

Model	Response Variable	Explanatory Variables	Residual Deviance	Residual DF	AIC	BIC
26	Status	Age + Agent	310.62	396	316.62	328.59
29	Status	Branch + Agent	211.91	396	217.91	229.88
31	Status	Direct + Agent	307.37	396	313.37	325.34
33	Status	Agent + Salary	316.84	396	322.84	334.81

Model 29 provides the new assessment criteria levels for the three-predictor model set with the R-Squared and accuracy values shown in Table 5.1.

Table 5.1 – Maximum R-Squared and Accuracy for Two-Predictor Models
(Extract of Table 5 in Appendix)

Explanatory Variables	R-Squared	Accuracy
Branch + Agent	0.510935	0.9223058

4.4. Three-Predictor Models

The 4 two-predictor models still under consideration, all with the Agent variable in common, create 6 three-predictor model combinations. Only one model reduces the AIC and BIC, raises the R-Squared and maintains the previous accuracy rate. No other models meet at least three of the four assessment criteria levels, the full details of which can be seen in Tables 6 and 7 in the appendix.

Using the final three-predictor model (37), all variations of interaction terms are added to determine if an interaction between any of the two variables is significant. With three variables, there are three different possible interaction terms, the details of which are shown in Table 8 below.

Table 8 – Final Model and Final Model with Interaction Terms
(Extract of Table 8 in Appendix)

Model	Response Variable	Explanatory Variables	Residual Deviance	Residual DF	AIC	BIC
37	Status	Agent + Branch + Direct	205.54	395	213.54	229.50
40	Status	Agent * Branch + Direct	177.03	394	187.03	206.98
41	Status	Agent + Branch * Direct	203.12	394	213.12	233.06
42	Status	Agent * Direct + Branch	205.46	394	215.46	235.40

Only two of the resulting models, 41 and 42, have a decrease in AIC value while only one of those two, 41, also has a decrease in BIC value. In addition, there is a marked improvement in R-Squared for model 41 which can be seen in Table 9 in the appendix.

At first glance, model 41 appears to be the best fit, but given that all three variables are categorical, the difference of deviance is analysed and the results can be seen in Table 10. As only one model shows a p-value that is significant at the 0.05 significance level, it confirms that model 41 is the best fit for the bank data.

Table 10 – Difference of Deviance Test for Final Model and its Nested Forms

Model	Difference in Residual Deviance	Difference in Residual DF	P-Value	Significant
Model 37 – Model 41	28.51	1	<0.00001	Yes
Model 37 – Model 42	2.42	1	0.119795	No
Model 37 – Model 43	0.08	1	0.777297	No

4.5. Prediction Ability

To ensure that model 41 is useful for prediction as well, the model is tested using a fraction of the dataset before calculating the prediction accuracy. To achieve this, the data are split into training and testing sets using a 70/30 ratio, respectively. The final model is fitted using the training set data and a classification table is created to determine the model accuracy. The fitted model is then used with the testing set to determine prediction accuracy, once again using a classification table, an example of which is shown in Table 11.1.

Table 11.1 – Classification Table for Test Set 1

Test Set 1	0	1
0	21	7
1	0	92

For testing set 1, the prediction accuracy equates to (21 + 92)/(21+92+7) = 0.941667. This process is completed nine more times, each with new random training and testing sets. The average of both the ten training sets and testing sets are taken and used as the model accuracy and prediction rates.

Table 13 – Accuracy Rates for the Final Model with Different Datasets
(Information extracted from Tables 9, 11 and 12 in Appendix)

Model Accuracy (Full Dataset)	Average Model Accuracy (Training Set)	Average Prediction Accuracy (Testing Set)
0.9223	0.9219	0.9200

As shown in Table 13, the difference between the model accuracy with the full dataset and the average model accuracy with the training set is only 0.0004. This shows that the final model selected produces virtually the same accuracy even when created using a fraction of the original data. In addition, the average prediction accuracy differs from the full dataset and training set model accuracies by only 0.0023 and 0.0019, respectively, which illustrates the consistent nature of the final model.

5. RESULTS

The final three-predictor model takes the form

$$\text{logit}(u) = 5.0347 - 1.2949*AN + 2.2576*BN - 2.3721*DN - 7.0344*AN:BN$$

where AN = Agent No, BN = Branch No, DN = Direct No, AN:BN = Agent No:Branch No, logit(u) = log-odds of a successful outcome and u = probability of a successful outcome.

The model shows that a customer

- lacking an agent will decrease the log-odds by 1.2949
- not being a branch user will increase the log-odds by 2.2576
- not being a direct user will decrease the log-odds by 2.3721
- lacking an agent and not being a branch user will decrease the log-odds by 7.0344.

Therefore, the probability of a random customer staying with the bank at the end of the 'lock-in' period depending on all combinations of personal circumstances is shown in Table 14 below.

Table 14 – Probability of Retention for all Combinations of the Final Model
(Calculations available in Table 15 in Appendix)

Probability	Agent	Branch User	Direct User	Agent:Branch User
99.32%	Yes	Yes	Yes	Yes
93.50%	Yes	Yes	No	Yes
99.93%	Yes	No	Yes	Yes
99.30%	Yes	No	No	Yes
97.68%	No	Yes	Yes	Yes
79.71%	No	Yes	No	Yes
26.16%	No	No	Yes	No
3.20%	No	No	No	No

Therefore, this model can be used for prediction purposes in the future, with a prediction accuracy rate of 92%, for the bank to determine the likelihood of a customer's decision to stay or leave the bank.

6. CONCLUSION

According to the results of this analysis, it is in the bank's best interests to place more importance on customers having a relationship with the bank as well as initiating specific points of contact. Having an agent is the best option with a 99.30% probability of retention, while the combinations of agent/direct user and agent/branch/direct user, are the 2nd and 3rd options available with probabilities of retention of 99.93% and 99.32%, respectively. However, if a customer doesn't want an agent, the next best option is for the customer to bank by both branch and telephone which only decreases the probability of retention by 1.62%. In conclusion, the goal for the bank is to direct customers to adopt the combination of having an agent and banking directly by telephone as this results in the highest likelihood of customer retention.

7. REFERENCES

Agresti, A. (2015) *Foundations of Linear and Generalized Linear Models*. New Jersey: John Wiley & Sons, Incorporated. Available at: http://ebookcentral.proquest.com.ezproxy3.lib.le.ac.uk/lib/leicester/reader.action?docID=1895564 (Accessed: 26 April 2017).

Alice, M. (2015) 'How to Perform a Logistic Regression in R', *R-Bloggers*, 13 Sept. Available at: https://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/ (Accessed: 21 April 2017).

Andale. (2015) 'Akaike's Information Criterion: Definition, Formulas', *Statistics How To*, 7 Sept. Available at: http://www.statisticshowto.com/akaikes-information-criterion/ (Accessed 24 April 2017).

Analytics Vidhya (2015) *Simple Guide to Logistic Regression*. Available at: https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/ (Accessed: 23 April 2017).

Frost, J. (2013) 'Multiple Regression Analysis: Use Adjusted R-Squared and Predicted R-Squared to Include the Correct Number of Variables', *The Minitab Blog*, 13 June. Available at: http://blog.minitab.com/blog/adventures-in-statistics-2/multiple-regession-analysis-use-adjusted-r-squared-and-predicted-r-squared-to-include-the-correct-number-of-variables (Accessed: 5 May 2017).

Frost, J. (2015) 'The Danger of Overfitting Regression Models', *The Minitab Blog*, 3 Sept. Available at: http://blog.minitab.com/blog/adventures-in-statistics-2/the-danger-of-overfitting-regression-models (Accessed: 1 May 2017).

The Methodology Centre (2017) *AIC vs. BIC*. Available at: https://methodology.psu.edu/node/504 (Accessed: 3 May 2017).

Penn State Eberly College of Science (2017) *STAT 504: Analysis of Discrete Data.* Available at: https://onlinecourses.science.psu.edu/stat504/node/49 (Accessed: 20 April 2017).

Rodriguez, G. (2017) '5 Generalized Linear Models', *Introducing R*. Available at: http://data.princeton.edu/R/glms.html (Accessed: 20 April 2017).

Wagenmakers, E. and S. Farrell. (2004) 'AIC Model Selection Using Akaike Weights', Psychonomic Bulletin & Review, 11(1), pp. 192-196. Available at: https://www.researchgate.net/publication/8588301_AIC_model_selection_using_Akaike_weights (Accessed: 23 April 2017).

8. APPENDIX

Table 2 – Single Predictor Models with AIC and BIC

Model	Response Variable	Explanatory Variable	Residual Deviance	Residual DF	AIC	BIC
0	Status	Null	433.29	398	435.29	439.28
1	Status	Holder	433.29	397	437.29	445.27
2	Status	Post	17.32	204	407.32	1185.16
3	Status	Type	422.37	395	430.37	446.32
4	Status	Amount	432.96	397	436.96	444.93
5	Status	Name	278.80	389	298.80	338.69
6	Status	Code	280.01	389	300.01	310.50
7	Status	Balance	431.07	397	435.07	443.04
8	Status	Term	425.46	397	429.46	437.44
9	Status	Age	423.07	397	427.07	435.05
10	Status	Region	402.12	389	422.12	462.01
11	Status	Membership	429.38	397	433.38	441.36
12	Status	Gender	433.29	397	437.29	445.27
13	Status	Branch	420.06	397	424.06	432.04
14	Status	Direct	426.50	397	430.50	438.48
15	Status	Internet	431.69	397	435.69	443.67
16	Status	Agent	319.59	397	323.59	331.57
17	Status	Acron	315.57	356	401.57	573.10
18	Status	Salary	425.87	397	429.87	437.85

Table 3 – Single Predictor Models with R-Squared and Accuracy

Model	Response Variable	Explanatory	R-Squared	Accuracy
1	Status	Holder	7.57E-09	0.7669173
2	Status	Post	0.96004	0.9874687
3	Status	Type	0.02522	0.7669173
4	Status	Amount	7.80E-04	0.7669173
5	Status	Name	0.35655	0.8195489
6	Status	Code	0.31105	0.8145363
7	Status	Balance	5.14E-03	0.7669173
8	Status	Term	0.01808	0.7669173
9	Status	Age	0.02360	0.7619048
10	Status	Region	0.07194	0.7744361
11	Status	Membership	9.03E-03	0.7669173
12	Status	Gender	2.93E-08	0.7669173
13	Status	Branch	0.03055	0.7669173
14	Status	Direct	0.01568	0.7669173
15	Status	Internet	3.69E-03	0.7669173
16	Status	Agent	0.26241	0.7669173
17	Status	Acron	0.27169	0.7994987
18	Status	Salary	0.01713	0.7593985

Table 4 – Two-Predictor Models with AIC and BIC

Model	Response Variable	Explanatory Variables	Residual Deviance	Residual DF	AIC	BIC
19	Status	Term + Age	408.68	396	414.68	426.64
20	Status	Term + Branch	407.69	396	413.69	425.66
21	Status	Term + Direct	419.22	396	425.22	437.18
22	Status	Term + Agent	319.12	396	325.12	337.09
23	Status	Term + Salary	420.14	396	426.14	438.11
24	Status	Age + Branch	410.48	396	416.48	428.44
25	Status	Age + Direct	416.66	396	422.66	434.63
26	Status	Age + Agent	310.62	396	316.62	328.59
27	Status	Age + Salary	416.78	396	422.78	434.74
28	Status	Branch + Direct	414.20	396	420.20	432.16
29	Status	Branch + Agent	211.91	396	217.91	229.88
30	Status	Branch + Salary	408.43	396	414.43	426.39
31	Status	Direct + Agent	307.37	396	313.37	325.34
32	Status	Direct + Salary	416.75	396	422.75	434.72
33	Status	Agent + Salary	316.84	396	322.84	334.81

Table 5 – Two-Predictor Models with R-Squared and Accuracy

Model	Response Variable	Explanatory Variables	R-Squared	Accuracy
19	Status	Term + Agent	0.0568130	0.7619048
20	Status	Term + Branch	0.0590906	0.7669173
21	Status	Term + Direct	0.0324904	0.7669173
22	Status	Term + Agent	0.2634941	0.7669173
23	Status	Term + Salary	0.0303504	0.7619048
24	Status	Age + Branch	0.0526580	0.7669173
25	Status	Age + Direct	0.0383862	0.7619048
26	Status	Age + Agent	0.2831123	0.7919799
27	Status	Age + Salary	0.0381189	0.7568922
28	Status	Branch + Direct	0.0440760	0.7669173
29	Status	Branch + Agent	0.5109350	0.9223058
30	Status	Branch + Salary	0.0573895	0.7744361
31	Status	Direct + Agent	0.2906162	0.7669173
32	Status	Direct + Salary	0.0381819	0.7568922
33	Status	Agent + Salary	0.2687572	0.7969925

Table 6 – Three-Predictor Models with AIC and BIC

Model	Response Variable	Explanatory Variables	Residual Deviance	Residual DF	AIC	BIC
34	Status	Agent + Age + Branch	211.42	395	219.42	235.37
35	Status	Agent + Age + Direct	298.36	395	306.36	322.32
36	Status	Agent + Age + Salary	308.43	395	316.43	332.38
37	Status	Agent + Branch + Direct	205.54	395	213.54	229.50
38	Status	Agent + Branch + Salary	215.57	395	215.57	231.52
39	Status	Agent + Direct + Salary	302.81	395	310.81	326.76

Table 7 – Three-Predictor Models with R-Squared and Accuracy

Model	Response Variable	Explanatory Variables	R-Squared	Accuracy
34	Status	Agent + Age + Branch	0.5120676	0.9223058
35	Status	Agent + Age + Direct	0.3114104	0.8145363
36	Status	Agent + Age + Salary	0.2881796	0.7869674
37	Status	Agent + Branch + Direct	0.5256275	0.9223058
38	Status	Agent + Branch + Salary	0.5209564	0.9172932
39	Status	Agent + Direct + Salary	0.3011532	0.7619048

Table 8 – Three-Predictor Interaction Models with AIC and BIC

Model	Response Variable	Explanatory Variables	Residual Deviance	Residual DF	AIC	BIC
37	Status	Agent + Branch + Direct	205.54	395	213.54	229.50
40	Status	Agent * Branch + Direct	177.03	394	187.03	206.98
41	Status	Agent + Branch * Direct	203.12	394	213.12	233.06
42	Status	Agent * Direct + Branch	205.46	394	215.46	235.40

Table 9 – Three-Predictor Interaction Models with R-Squared and Accuracy

Model	Response	Explanatory Variables	R-Squared	Accuracy
37	Status	Agent + Branch + Direct	0.5256275	0.9223058
40	Status	Agent * Branch + Direct	0.5914288	0.9223058
41	Status	Agent + Branch * Direct	0.5312300	0.9223058
42	Status	Agent * Direct + Branch	0.5258288	0.9223058

Table 11 – Testing Set Accuracy Rates

Model	Test 1	Test 2	Test 3	Test 4	Test 5
Agent + Branch + Direct	0.9417	0.9333	0.9333	0.9083	0.9083
Model	Test 6	Test 7	Test 8	Test 9	Test 10
Agent + Branch + Direct	0.9167	0.9500	0.8750	0.9000	0.9333

Table 12 – Training Set Accuracy Rates

Model	Train 1	Train 2	Train 3	Train 4	Train 5
Agent + Branch + Direct	0.9140	0.9176	0.9104	0.9283	0.9283
Model	Train 6	Train 7	Train 8	Train 9	Train 10
Agent + Branch + Direct	0.9247	0.9104	0.9427	0.9247	0.9176

Table 15 – Final Model Probability Calculations

$$\begin{matrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{matrix} * \begin{matrix} 5.0347 \\ -1.2949 \\ 2.2576 \\ -2.3721 \\ -7.0344 \end{matrix} = \begin{matrix} 5.035 \\ 2.663 \\ 7.292 \\ 4.920 \\ 3.740 \\ 1.368 \\ -1.037 \\ -3.409 \end{matrix}$$