To accomplish this task, we decided to apply the regression logistics algorithm. Logistic regression algorithm is commonly used in predicting binary outcomes. In the case of churning, we want to predict if a new customer or existing customers are likely to stay or exit. We will examine a dataset composed of 10,000 records. ‘exited’ column in this dataset is regarded as the dependent variable which is the subject of this analysis, and the rest of the variables are the regressors or independent variables.

View Data: 
10,000 rows 
Customer Id 
15634602 
15647311 
15619304 
15701354 
15737888 
15574012 
15592531 
15656148 
15792365 
15592389 
15767821 
15737173 
Churn Modelling (P12-Churn Modelling) 
Show aliases 
x 
Dataset 
Exi ted 
Exi ted 
Stayed 
Exi ted 
Stayed 
Stayed 
Exited 
Stayed 
Exi ted 
Stayed 
Stayed 
Stayed 
Stayed 
Gender 
Female 
F emale 
F emale 
Female 
F emale 
Male 
Male 
F emale 
Male 
Male 
Male 
Male 
Geogr aph y 
France 
Spain 
Fr ance 
France 
Spain 
Spain 
Fr ance 
German y 
Fr ance 
Fr ance 
Fr ance 
Spain 
Has Cr Card 
Is Active Member 
Num Of Products 
Ron Number 
Surname 
Hargr ave 
5 Mitchell 
Bartlett 
10 
Il 
Bear ce 
Andrews 
12 
Age 
41 
42 
43 
44 
29 
44 
27 
31 
Balance 
0.00 
83,807.86 
159,660.80 
0.00 
125,510.82 
113,755.78 
0.00 
115,046.74 
142,051.07 
134,603.88 
102,016.72 
0.00 
Credit Score 
Estima ted Salar y 
101,348.88 
112,542.58 
113,931.57 
93,826.63 
79,084.10 
149,756.71 
10,062.80 
119,346.88 
74, 50 
71,725.73 
80,181.12 
76,390.01 
Number of Records 
Tenure 

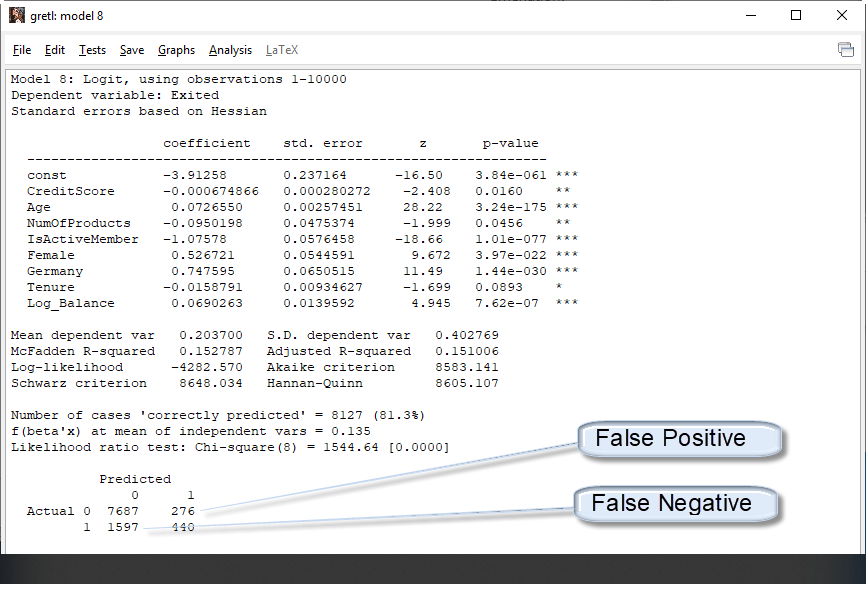
Tableau showed correlation with ‘Has Credit Card’ and ‘ ‘IsActive’ at 20% stayed/exited reference line.

Pages 
Filters 
Ma rks 
000 Automatic 
15.5%- 17.5% 
iii Columns 
Rows 
Gender 
Gender 
SUM(Number of Rec„ 
Actuals 
Gender 
Evan's Awesome AB Tools (home): 
Sample Size Calculator Chi-Squared Test Sequential Sampling 2 Sample T-Test Survival Times 
Question: Does the rate of success differ across two groups? 
Count Data 
z 
ssoo 
sooo 
450 
4000 
3500 
3000 
2500 
2000 
1500 
1000 
Sam le 
pie 2: 
# successes 
898 
# trials 
4543 
23.8% 
457 
Verdict: 
Confidence interval 
26.4% 
[ clear ] [ ] 
Color 
Detail 
Size 
Tooltip 
Label 
Exited 
a Exited 
a SLIM(Number 
Exited 
• Stayed 
• Exited 
St aye C 
4,559 
Stayed 
3,404 
Sample I is more successful 
Confidence level: 
If the experiment is re 
the reported coffidence intenai_ 
will fail within 
It is also the percent of the time no difference will be detected between the two groups, assuming no difference exists. 
If you like this, check out Wizard the easy Mac statistics app. 
Female 
898 
Male 

After performing data mining in Tableau, we built a logistic regression model in Gretl and performed 5 backward eliminations. We made dummy variables ‘Spain’, Germany’ and ‘France’ from the ‘Geography’ variable and ‘Male’ and ‘Female’ variables from ‘Gender’ variable. We included ‘Female’, ‘Spain’ and ‘Germany’ in the model along with the other independent variables in the dataset. In each run of the model, Gretl recommended removal of a variable. Table below shows the summary of each elimination. The main criteria for keeping a variable in the model was that the p-value to be below out threshold of 0.5 for the variables and the Adjusted R-Squared increasing for each model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| BWElimination Number | Variable eliminated | Variable P-Value | Model’s Adjusted R-squared before/after removal | Adjusted R-Squared Difference |
| 1 | Spain | 0.6181 | 0.150787/ 0.150961 | 0.000174 |
| 2 | HasCrCard | 0.4489 | 0.150961/ 0.151102 | 0.000141 |
| 3 | EstimatedSalary | 0. 3091 | 0.151102/ 0.151197 | 0.000095 |
| 4 | Tenure | 0.0873 | 0.151197/ 0.151106 | -0.000091 |

As shown in elimination 4 ‘Tenure’ was removed, but not by recommendation from Gretl, but because we wanted to see the impact of removal to test the p-value threshold. It shows that the Adjusted R-Squared was not impacted by much, so we reincluded ‘Tenure’ in the model. After transforming the ‘Balance’ variable to Log10(Balance +1) for better uniformity, we got the result shown below.



From the confusion Matrix, the accuracy and error rates we calculated as shown below.

Accuracy Rate = Correct/Total = (7687+440)/10000 = 81.27 %

Error Rate = Wrong/Total = (276+1597)/10000 = 18.73 %

Based on the current analysis, we deem the accuracy rate to be not satisfactory. We look to maximize this rate by introducing new variables from our data mine.