Churn Modelling:

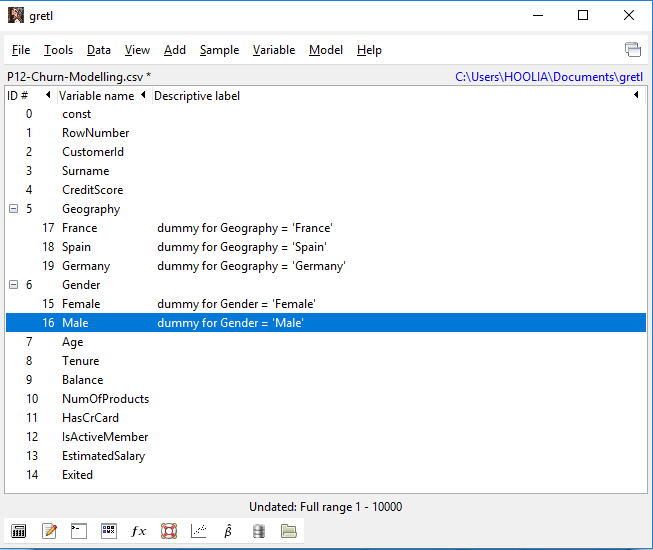
Problem Statement: Unusual churn rates at a Bank operating in 3 countries (France Spain, Germany), assess and address the problem by providing insights from data set. CustomerID, Surname, Credit Score, Geography, Gender, Age, tenure, Products Involved in, Active member, Estimated Salary, Exited (0-stayed, 1-exited)

Questions answered by this model:

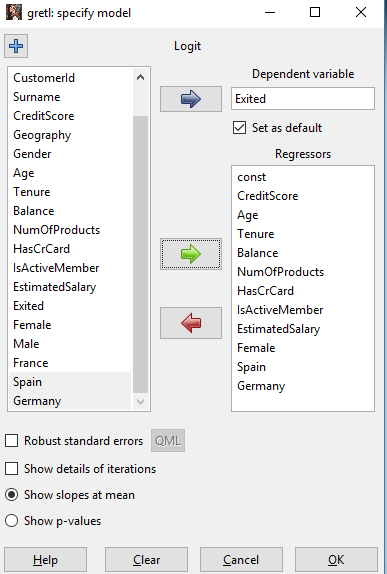
* Should the person be given loan?
* Identify transactions fraudulent?
* Should they be Approved for Credit Card?

Prior experience helps build model to handle churn rate based on factors (independent variables) that influence the outcome.

Dummy variables are named.

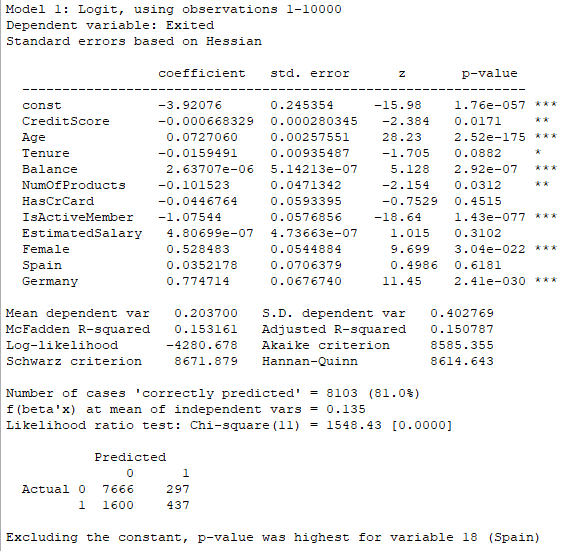


Modelling :

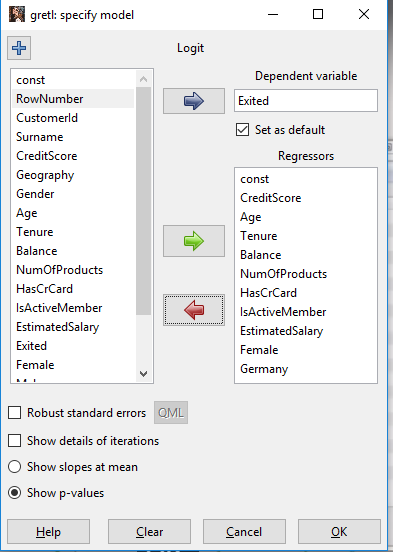


* In gender dummy variables, we choose Female as baseline. 0 is male, 1 female.
* In Country we choose France is baseline.

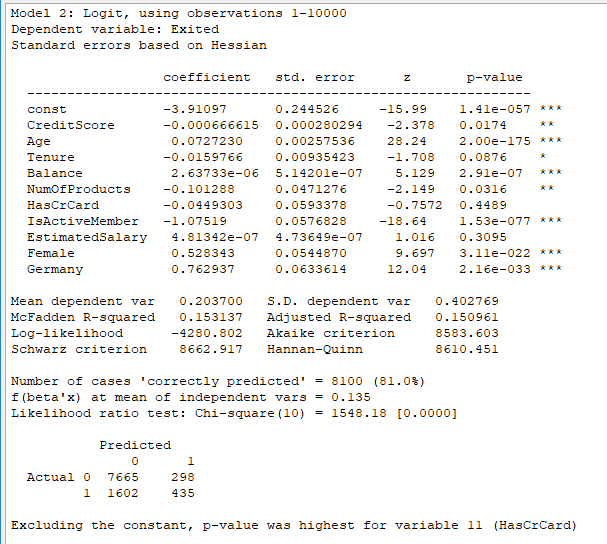
Logit Model:



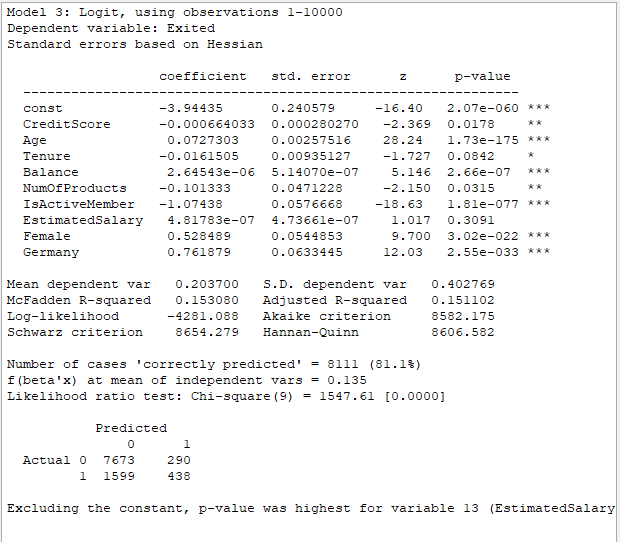
* Eliminate variables crossing p>5%, iterate for next highest p value.
* By observing the p-values, we see people in Germany are substantially different in terms of their probability to leave the bank to people who live in France.
* People in Spain are as likely to leave the bank as people in France, hence not a significant contributor to the model. For example, if we take three people person from Spain and France will exhibit same behavior whereas Germany would be the higher likelihood to leave the bank (coefficient is 0.77). Another reason could be a competitor who taking away people in Germany, or customer service in Germany is not as good or product released in Germany is not doing good unlike Spain and France.



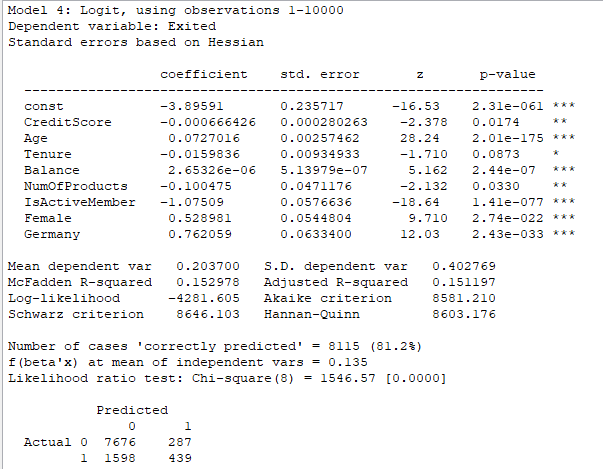
* Spain removed since it crossed p value threshold



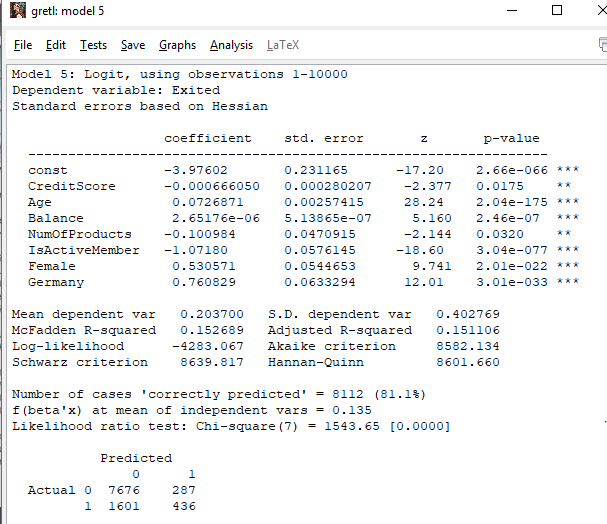
* Remove HasCrCard, exceeds p value of 5%. It means people with or without credit card while holding other parameters same, will have same likelihood of leaving the bank.



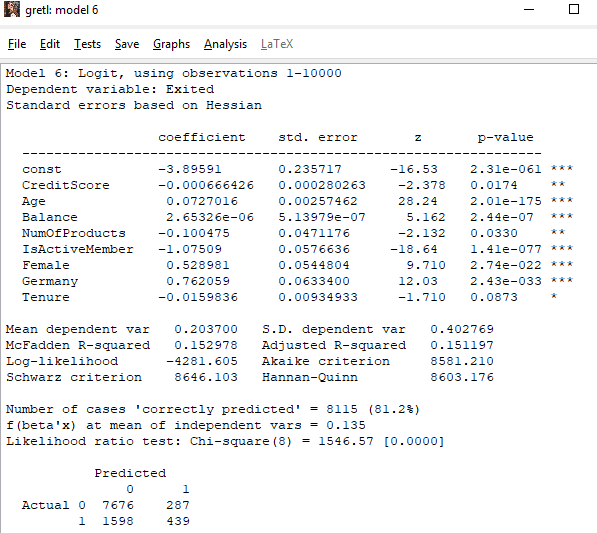
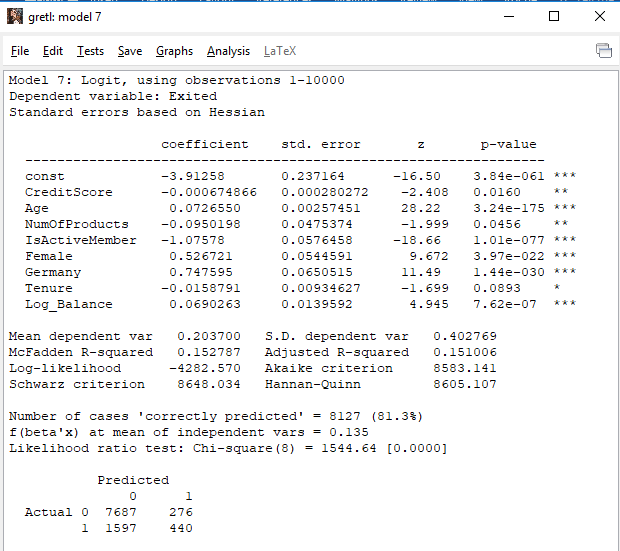
* Remove Estimated salary, exceeds p value of 5%. It means salary not a significant factor parameter

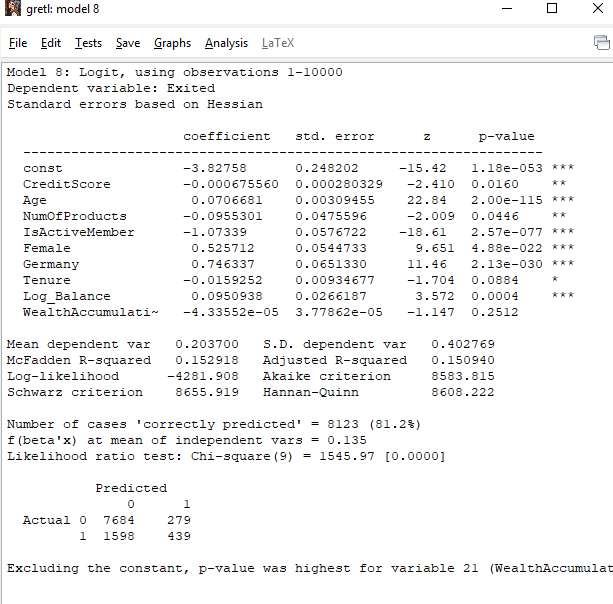


* We now remove the Tenure variable since it exceeds the p value over 5%.

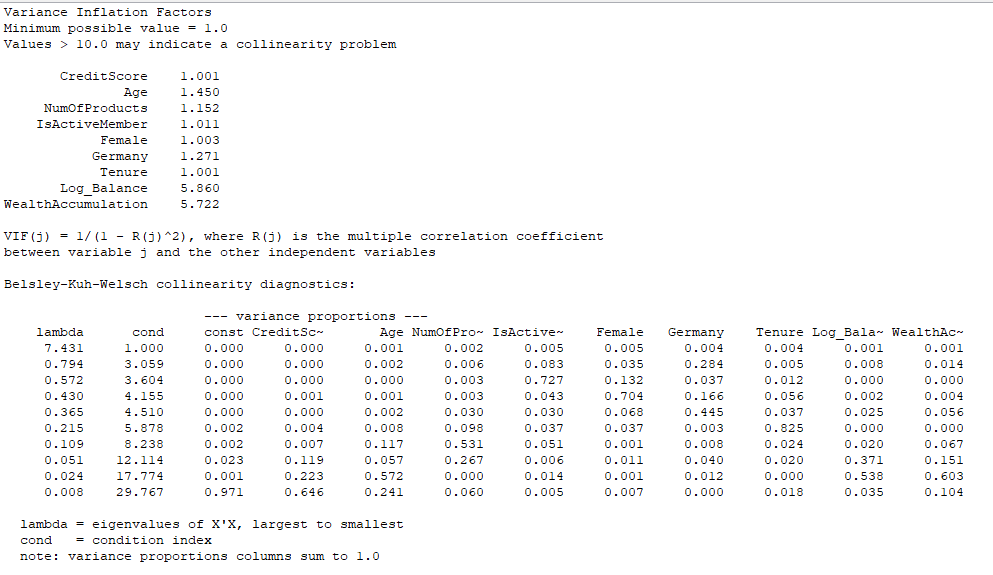


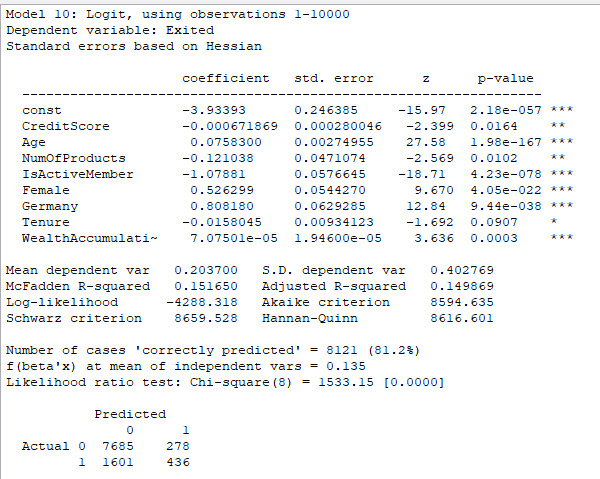
* R square and accuracy goes down by negligible. Tenure is generally a good factor, but in this case it wasn’t. Howeve, Tenure generally plays significant role and will add the IV back to model.

 After applying Transaformtaions to Balance, LogBalance=log10(balance+1), we see accuracy improves. But R squared goes down. Greater the starting balance. Lesser the 1 unit significance. In this case it is 10% increase. Its consistent over previous method.

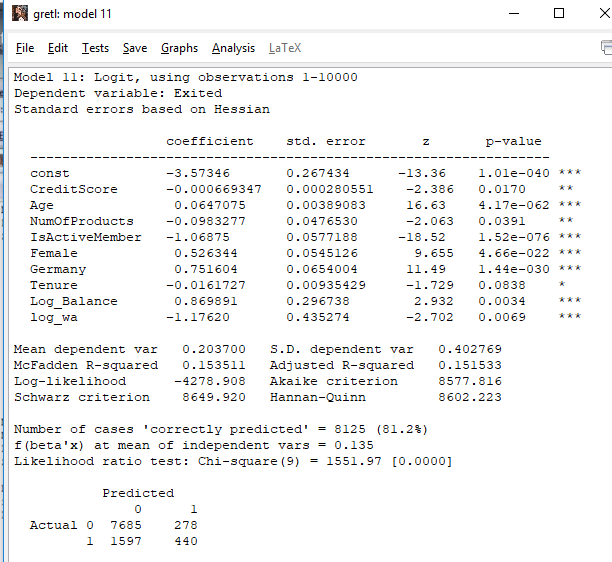
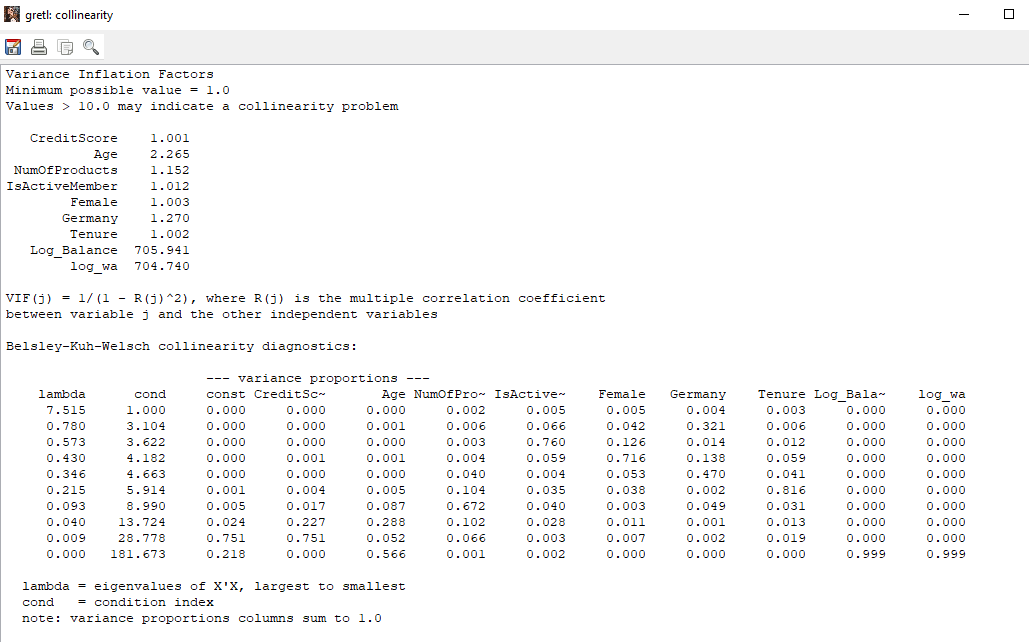
New variable created, Wealth accumulation which divides balance by age, but has no significant effect on the model. The newly created variable is collinear with Age, Balance, Log Balance.

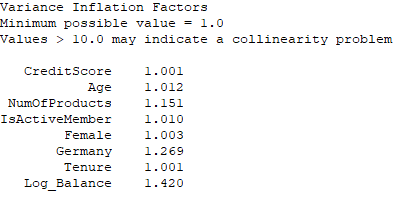
Multi-collinearity using VIF:

LogBalance and WealthAccumulation have collinearity of 5.8 and 5.7. Hence, remove LogBalance

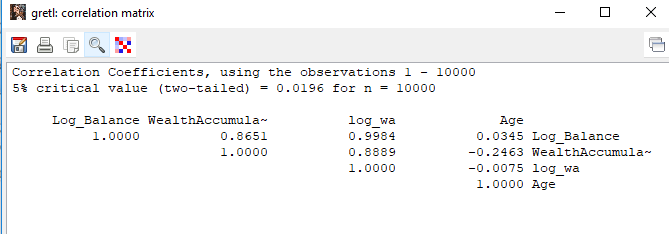
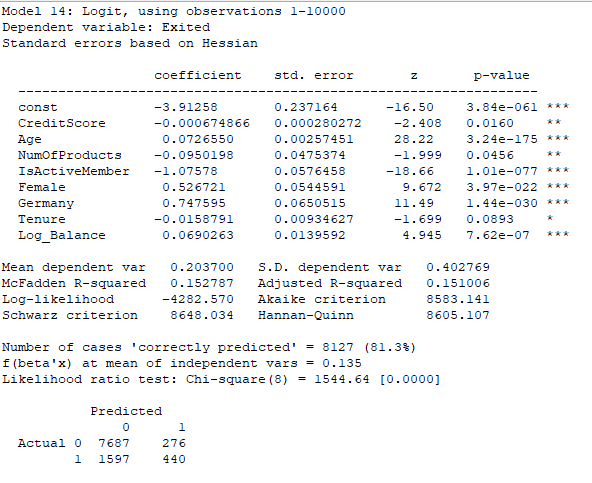


We notice, Wealth Accumulation plays significant role. LogBalance and WealthAccumulation have high collinearity. Hence, removing one of them makes huge difference.

New variable log\_wa = log10(balance/Age +1) is created. Log\_balance and Log\_WA has significant relation with the dependent variable (Exited). When we run collinearity test, we see higher values for these variables meaning they are highly collinear. Remove log\_WA.



Using VIF we were able to measure the degree of collinearity wrt to IV and observe changes in Rsquare and accuracy of the model.

Understanding the correlation between variables, excluding variables with highest correlation produces robust model. When two variables are mutually correlated, its hard to study the impact it has on the dependent variable. Since the correlation affects others Independent variable also to change. Thus we need to break multi-collinearity since it breaks model.