|  |
| --- |
| Joseph Deery  40078793  Contents  [1.0 Introduction and Background 1](#_Toc58421860)  [2.0 Methodology 1](#_Toc58421861)  [3.0 Results and Discussion 2](#_Toc58421862)  [4.0 Conclusions 14](#_Toc58421863)  [5.0 References 15](#_Toc58421864)  [6.0 Appendix – R code 15](#_Toc58421865) |

|  |
| --- |
| QueensBank |
| Exploration and Analysis of Factors Relating to Term Deposit Subscriptions |
| A Data-Driven Approach to Marketing at QueensBank |

# Introduction and Background

The ever-growing landscape of direct consumer marketing has gradually diminished its impact and effectiveness over time (Moro, Cortez and Loreano, 2011). This, along with increasing industry competition, has compelled managers to adopt a highly selective approach to direct marketing, to increase its effectiveness and return on investment. This process can be enhanced using Business Intelligence, Data Mining and Analytics. In the past, banks have had two options when it comes to marketing: global, mass marketing, or direct. A study by Ling and Li (1998), found less than 1% of responses to mass campaigns were positive. Direct marketing has therefore often been the more desirable approach, targeting audiences with a greater likelihood of interest in a product or service, and, in theory, increasing the efficiency of the campaign (Ou et al. 2003). Despite this, some potential drawbacks, such as customer’s feeling privacy has been compromised (Page and Luding, 2003), are ever-more pertinent issues that banks should be wary of in a post GDPR climate. Predicting term-deposit subscriptions in banks through advanced analytics techniques has been a topic of research in recent years. A study by Colaianni, Magdangal and Mitchell (2016), using banking data, applied logistic regression and a binary decision tree to determine factors most correlated and predictive of term deposit descriptions. The study found that job type, whether a customer had defaulted prior, and previous loans had strong relationships with subscription decisions. Socio-Economic data, such as interest rates and employment numbers were also found to be predictive. This study adopts the CRISP-DM methodology to perform a similar exploration and analysis of QueensBank’s data, adapted from a previous telemarketing campaign. The analysis involves client data, social and economic data, and previous marketing data. The purpose is to explore the relationships of each variable for both inference and prediction on whether a customer will subscribe to a term deposit. The hypothesis for each variable will be that there is no significant relationship with the target variable (subscription), however it is expected that certain economic and client variables such as interest rate (euribor3m) and job, for example will reject this hypothesis and show significant relationships. These relationships will be explored through visualisation, correlation/association and regression, using Rstudio.

# Methodology (531)

* 1. *Data Quality/Preparation*

This step was hugely important in ensuring accuracy and quality of any analysis and to avoid a “garbage in, garbage out” scenario. The dataset was imported into RStudio using the “readxl” package. The data frame was then summarised to identify any errors, such as outliers, missing values, and formatting issues. Many of the categorical variables were converted to factors, as RStudio automatically stores these as “character” vectors, which are of little analytical use. Outliers were identified further using basic plots, visualisations, and descriptive functions. Extreme outliers were removed in 5 variables, “Age”, “Campaign”, “Duration”, “Previous” and “Cons\_Conf\_Indx”. These contextual outliers were a mix of viable/unviable, so some critical judgement was used. For example, maximum Age was reduced from 98 to 80, with 98 likely a viable customer age, however it was determined this would skew any visualisations or models produced. On the contrary, the maximum value for Consumer Confidence Index was 999, deviating significantly from the mean and likely to be an incorrect value. Missing data was largely left untouched, due to its potential use in predictive modelling, and could be removed on a case-by-case basis for visualisations.

* 1. Data Visualisation

Data Visualisations were produced using the ‘ggplot2’ package in Rstudio. The flexibility and customisation of this package allow for deeper and more aesthetic visualisations than base plots. Several key variables such as ‘job’, ‘education’, ‘Euribor3m’, ‘campaign’ and ‘cons\_conf\_idx’ were used to provide visual insights into the relationship with the dependent variable, ‘subscribed’, and five visualisations were produced. The best visualisations for categorical variables are often simple bar charts, however a density plot, scatterplot(jitter) and boxplot were also produced with extra variables included to show deeper relationships with customer subscriptions. The labelling was kept as minimal as possible, for ease of interpretation.

2.3 Correlation and Regression Analysis

This report uses Association/Correlation and Logistic Regression to provide both inference and predictive insights into factors related to the dependent binomial outcome variable. For correlation, the ‘chi-squared’ test was used to show the relationships between all categorical variables and the binomial output, while the ‘t test’ was applied to the independent numeric variables, the output of which are analysed and discussed in Section 3. Logistic regression was then used to create a model for inference and prediction. This method was chosen as the response variable is a binomial “yes/no”, and the output should show the significance each variable has on the odds ratio. Two preliminary models were created, and a final model (final\_model) including twelve independent variables that had the strongest correlations in previous tests was settled upon, as it best fit the data. Assumption checks and tests for predictive accuracy were carried out on the final model, shown, and interpreted in Section 3.

# Results and Discussion (1200)

* 1. Descriptive Statistics

Tables 1 and 2 show the descriptive summaries for both Numeric and Categorical variables respectively, post data quality fixes. Table 1 shows the skewness, distribution, mean and average of the data. Importantly, it shows effective removal of outliers for the most important variables with minimal standard deviation from the mean and skewness.

**Table 1**: Descriptive Statistics for Numeric Variables

Table 2 shows the descriptive statistics for categorical variables, having been correctly formatted as factors in R. It The table highlights the top categories for each variable as well as the frequency of each category. Giving a good overview of the spread of data across each variable.

**Table 2 :** Descriptive Statistics for Categorical Variables

|  |  |  |
| --- | --- | --- |
| Variable | Top Categories | Frequency |
| Job | Admin  Blue Collar  Technician  Services  Management  Retired  Other | 8326  7410  5381  3186  2347  1354  4946 |
| Marital Status | Married  Single  Divorced | 20012  9183  3679 |
| Education | Uni. Degree  High School  Basic 9y  Prof. Course  Basic 4y  Basic 6y  Other | 9710  7603  4828  4198  3337  1846  1428 |
| Default | No  Unknown  Yes | 26063  6884  3 |
| Housing Loan | No  Yes  Unknown | 27210  4937  781 |
| Personal Loan | No  Yes  Unknown | 27210  4937  781 |
| Contact | Cellular  Telephone | 290934  12016 |
| Month | May  Jul  Aug  Jun  Nov  Apr  Other | 10962  5759  4930  4271  3286  2110  1632 |
| Day of Week | Thu  Mon  Wed  Tue  Fri | 6919  6788  6919  6430  6251 |
| Previous Outcome | Nonexistent  Failure  Success | 28429  3418  1103 |
| Subscribed | No  Yes | 29238  3712 |

* 1. Visualisations

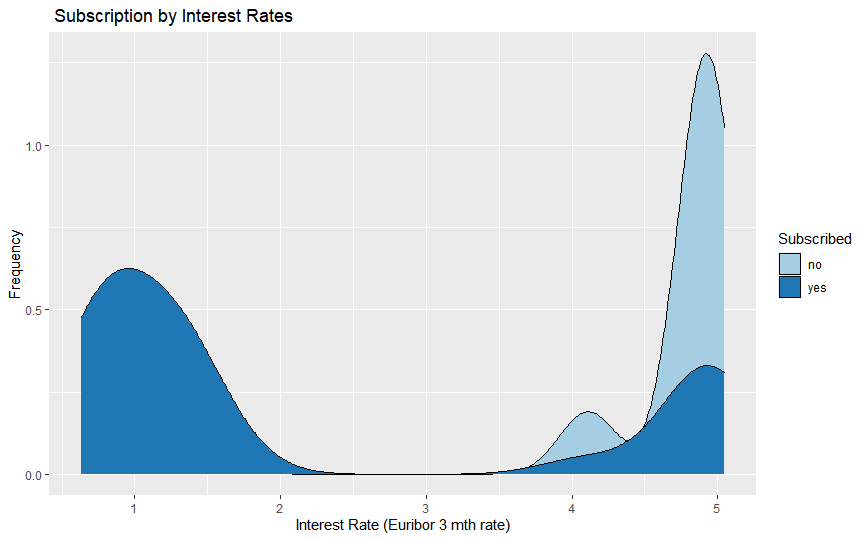
**Figure 1**: Density Plot: Subscriptions by Interest Rate

Figure 1 visualises the effect of the euribor3m interest rate on term deposit subscriptions. This indicator is based on the average bank interest rates within the Eurozone. In theory, as these increase, subscriptions should be positively influenced. This distribution of “yes” subscriptions across interest rates however appears to be bimodal in this case, with the dip occurring at the mid-level interest rates and spiking at the two extremes. This may be explained by another confounding variable that is influencing both the independent and dependent variables. The “no” responses also appear heavily left-skewed towards higher interest rates. This may be described by other external economic factors, or perhaps increased levels of direct marketing by QueensBank that have been that were received unfavourably by customers. The data does however show an expected rise in term deposit subscriptions towards higher interest rates.

**Figure 2**: Bar Chart: Term Deposit Subscriptions By month last Contacted

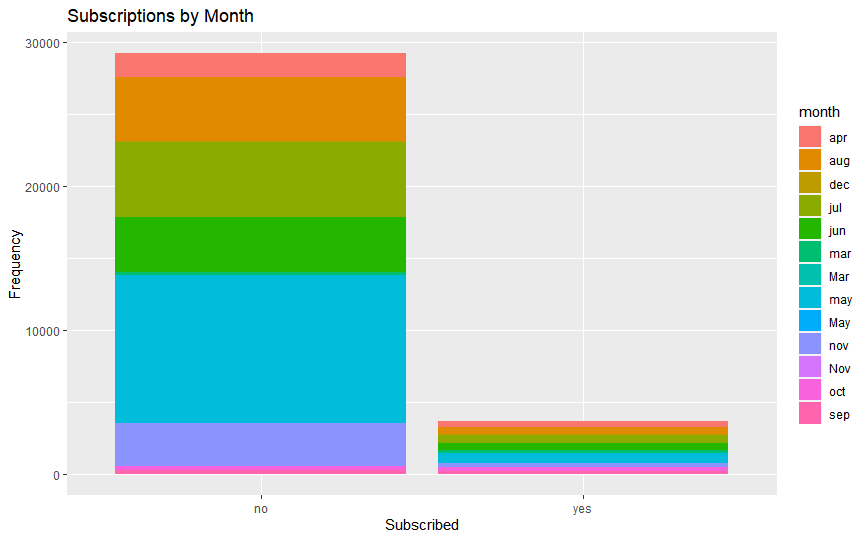


Figure 2 shows the distribution of yes/no subscriptions across months contacted during the last marketing campaign at QueensBank. This chart shows a both a high number of contacts and high number of “no” responses during the middle of the year, especially May. This is likely due to increased campaigning during the period. Interestingly however, the proportion of “yes” responses appears to be higher during the month of October and September. This may suggest customers are more receptive to marketing during these times of year, or when they are subjected to less sales calls altogether. This could also be a result of interest rates or other social/ economic factors, and a greater sample of data for these months would provide more clarity on whether the bank should increase marketing at this time, or whether they would see a greater ROI than over the summer period.

**Figure 3:** Scatterplot: Subscriptions by Previous Campaign Outcome and Previous Contact

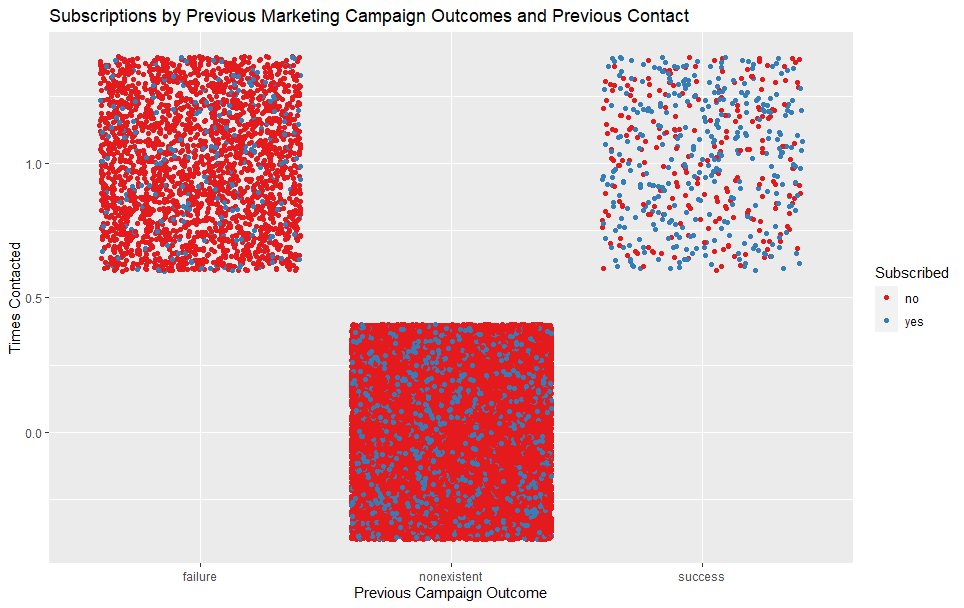
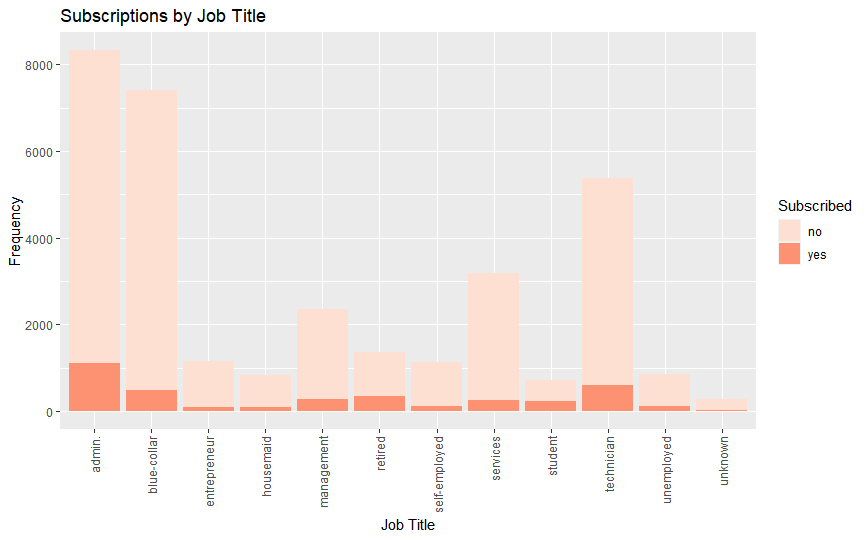
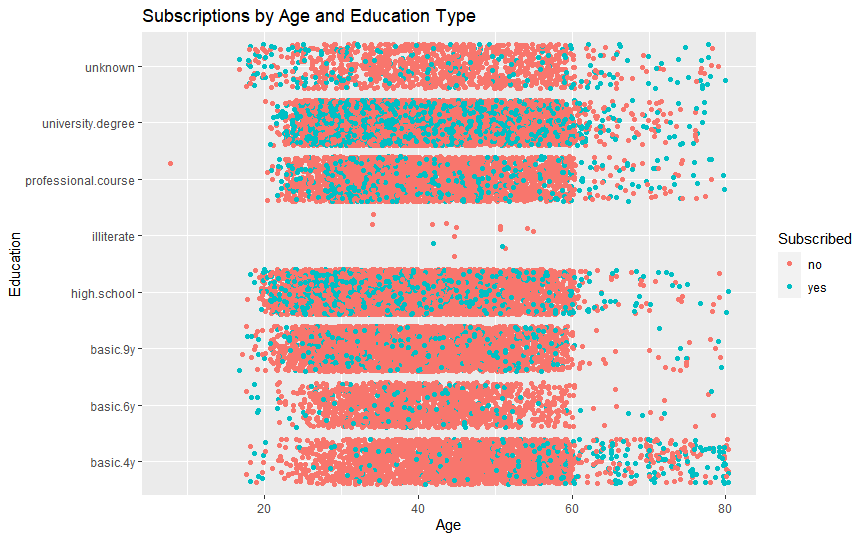


Figure 3 shows the effect previous contact with customers has on campaign success and subscriptions. Firstly, it shows that the “success” campaign was correctly evaluated, with a higher proportion than yes/no responses. It is also clear from this visualisation that previous contact is an important factor in determining subscriptions, showing a positive relationship with the decision to subscribe to a term deposit for those that have been contacted, compared to those that received no previous contact. This may be explained by trust and rapport being established between QueensBank and Client, or by customers simply having time to reflect and do research based on the initial call, enabling an informed decision to be made in follow-up calls. This visualisation shows the importance of keeping contact with clients, a strategy which should be adopted in future marketing campaigns. Figures 4 and 5 explore the relationships between client age, job and education backgrounds and the dependent variable. Both students and retired customers, based on the data, appear most receptive to marketing campaigns, with a higher proportion of “yes” responses than any other “job” type. The sample size for these roles however is limited compared to other job types and it is therefore unclear whether targeting these groups with more aggressive marketing would pay off, based on this visualisation alone.

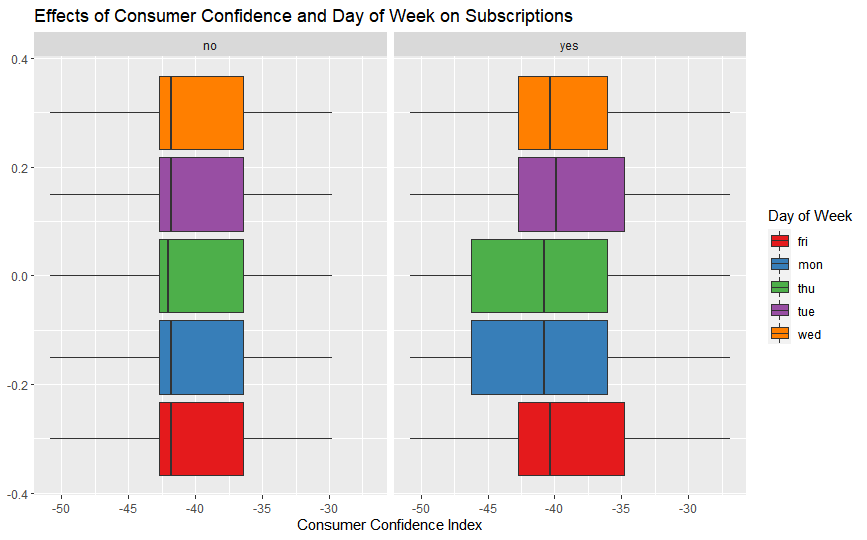
**Figure 4**: Subscriptions by Job Title

**Figure 5:** Subscriptions by Age and Education Type

**Figure 6:** Consumer Confidence and Day of Week on Subscriptions

Figure 6 explores the relationship between day of week and consumer confidence through the consumer confidence index levels, (The higher the index, the greater the level of consumer confidence). The boxplot shows the distribution of yes/no responses seems to be relatively even except on Mondays and Thursdays, with a higher number of “yes” responses than any other day of the week.

**Figure 6**: Boxplot: Effects of consumer confidence and day of week on Subscriptions



The plot also shows a higher average consumer confidence across each day within the “yes” response category. Consumer confidence seems to be higher on average on Tuesdays and Fridays, so it may be worth increasing marketing calls on these days, while QueensBank should also closely follow the consumer confidence index and adjust marketing levels accordingly, with a focus on targeted marketing where confidence levels are higher.

3.3 Correlation and Association

**Figure 7**: Pearson’s Chi-Squared Test Output for key Categorical Variables





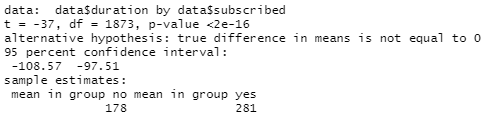


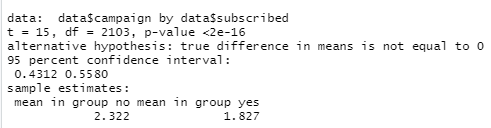


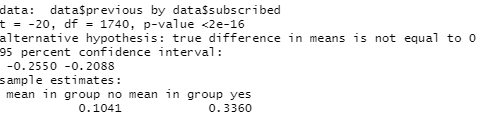


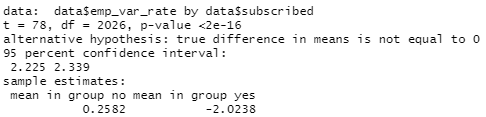


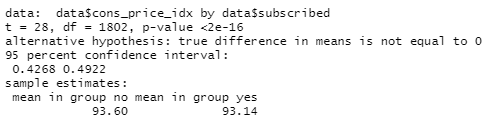
**Figure 8**: Welch T Test Output for key Numeric Correlations with binomial Dependent

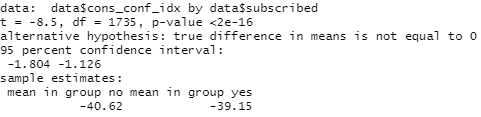


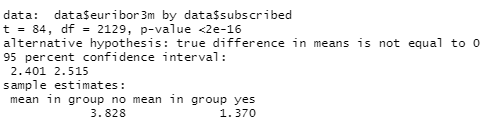


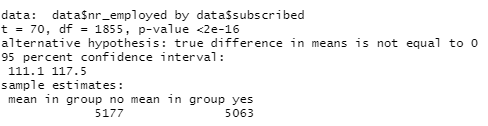












Figures 7 & 8 explore the measures of both correlation and association between key variables and the dependent “subscribed” variable, using the chi-squared test for categorical variables and the t.test for numeric independent variables, respectively. Each categorical variable tested shows a significant relationship at the 0.05 level when tested individually, rejecting the null hypothesis, as shown by the p-value. The x squared, or chi squared values for both “month” and “poutcome” show the highest probabilities that a relationship exists with the dependent variable Similarly, each numeric variable tested individually show significant relationships at the 0.05 level, with the “euribor3m” interest rate t-value showing the highest probability that this relationship exists and is not by chance. This information informed the basis for the logistic regression model.

3.4 Logistic Regression Model Interpretation

**Table 3**: Odds Ratios Derived from Logistic Regression Model

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Estimate** | **Odds Ratio** | **Std Error** | **Z Value** | **Pr(>|z|)** | **Sig.** |
| (Intercept) | 114.32103 | 4.46E+49 | 195.46306 | 0.58 | 0.5586 |  |
| Number Employed | -0.02582 | 3.20E-01 | 0.00237 | -10.88 | <2E-16 | \*\*\* |
| Consumer Conf. Idx | -0.04629 | 9.55E-01 | 0.0085 | -5.44 | 5.20E-08 | \*\*\* |
| Euribor3m Interest Rate | 1.26518 | 3.54E+00 | 0.20641 | 6.13 | 8.80E-10 | \*\*\* |
| Employee Var Rate | -1.13977 | 3.20E-01 | 0.1227 | -9.29 | <2E-16 | \*\*\* |
| Previous Outcome(None) | 0.48079 | 1.62E+00 | 0.08598 | 5.59 | 2.20E-08 | \*\*\* |
| Previous Outcome (Success) | 1.96953 | 7.17E+00 | 0.12521 | 15.73 | <2E-16 | \*\*\* |
| Campaign | -0.05608 | 9.46E-01 | 0.02328 | -2.41 | 0.016 | \* |
| Monday | -0.20442 | 8.15E-01 | 0.09863 | -2.07 | 0.0382 | \* |
| Thursday | 0.1701 | 1.19E+00 | 0.0956 | 1.78 | 0.0752 | . |
| Tuesday | 0.15876 | 1.17E+00 | 0.09887 | 1.61 | 0.1083 |  |
| Wednesday | 0.16122 | 1.18E+00 | 0.10176 | 1.58 | 0.1131 |  |
| August | -0.03006 | 9.70E-01 | 0.14174 | -0.21 | 0.832 |  |
| December | -0.42435 | 6.54E-01 | 0.26587 | -1.6 | 0.1105 |  |
| July | -0.20533 | 8.14E-01 | 0.15408 | -1.33 | 0.1826 |  |
| June | -0.32895 | 7.20E-01 | 0.15402 | -2.14 | 0.0327 | \* |
| March | 0.90242 | 2.47E+00 | 0.14252 | 6.33 | 2.40E-10 | \*\*\* |
| March | 0.99838 | 2.71E+00 | 1.42524 | 0.7 | 0.4836 |  |
| May | -1.18587 | 3.06E-01 | 0.10712 | -11.07 | <2E-16 | \*\*\* |
| May | 1.22539 | 3.41E+00 | 1.04449 | 1.17 | 0.2407 |  |
| November | -1.19776 | 3.02E-01 | 0.1988 | -6.02 | 1.70E-09 | \*\*\* |
| November | 0.97849 | 2.66E+00 | 1.16942 | 0.84 | 0.4027 |  |
| October | -0.86763 | 4.20E-01 | 0.2181 | -3.98 | 6.90E-05 | \*\*\* |
| September | -1.06801 | 3.44E-01 | 0.22144 | -4.82 | 1.40E-06 | \*\*\* |
| No Housing Loan | 9.24512 | 1.04E+04 | 195.12751 | 0.05 | 0.9622 |  |
| Unknown Housing Loan | 9.18012 | 9.70E+03 | 195.12761 | 0.05 | 0.9625 |  |
| Housing Loan | 9.10666 | 9.02E+03 | 195.12752 | 0.05 | 0.9628 |  |
| Telephone | -0.60348 | 5.47E-01 | 0.09999 | -6.04 | 1.60E-09 | \*\*\* |
| Basic 6y Education | 0.04162 | 1.04E+00 | 0.20784 | 0.2 | 0.8413 |  |
| Basic 9y Education | -0.0784 | 9.25E-01 | 0.15818 | -0.5 | 0.6202 |  |
| High School Ed | 0.09745 | 1.10E+00 | 0.14337 | 0.68 | 0.4967 |  |
| Illiterate | 0.57253 | 1.77E+00 | 1.17108 | 0.49 | 0.6249 |  |
| Professional Course | 0.248 | 1.28E+00 | 0.15466 | 1.6 | 0.1088 |  |
| University Degree | 0.23115 | 1.26E+00 | 0.14034 | 1.65 | 0.0995 | . |
| Education Unknown | 0.30601 | 1.36E+00 | 0.18048 | 1.7 | 0.09 | . |
| Blue-Collar | -0.55701 | 5.73E-01 | 0.1306 | -4.26 | 2.00E-05 | \*\*\* |
| Entrepreneur | -0.20573 | 8.14E-01 | 0.19298 | -1.07 | 0.2864 |  |
| Maid | -0.12626 | 8.81E-01 | 0.22782 | -0.55 | 0.5794 |  |
| Manager | -0.22086 | 8.02E-01 | 0.12539 | -1.76 | 0.0782 | . |
| Retired | 0.21192 | 1.24E+00 | 0.13024 | 1.63 | 0.1037 |  |
| Self Employed | -0.15439 | 8.57E-01 | 0.16724 | -0.92 | 0.3559 |  |
| Services | -0.42508 | 6.54E-01 | 0.14183 | -3 | 0.0027 | \*\* |
| Student | 0.23862 | 1.27E+00 | 0.14199 | 1.68 | 0.0928 | . |
| Technician | -0.15943 | 8.53E-01 | 0.10669 | -1.49 | 0.1351 |  |
| Unemployed | 0.13739 | 1.15E+00 | 0.1649 | 0.83 | 0.4047 |  |
| Job Unknown | -0.04518 | 9.56E-01 | 0.3432 | -0.13 | 0.8953 |  |

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 3 includes the summarised output and statistics of the logistic regression model, based on the key variables associated with deposit subscriptions at QueensBank. As outlined in the methodology, this model (“final\_model”) was the best from two preliminary models. It was determined to be the best fitting model, as it had the lowest AIC score of the three. The key things to focus on with this table are the estimate coefficient, odds ratio, and p-value. Interpreting the coefficients, the number employed (nr\_employed) variable is statistically significant at the 0.001 level and shows a negative relationship with the dependent variable; that is the less people employed, the more likely the probability that subscriptions to term deposits increase. Consumer Confidence Index and employee variable rate also show negative relationships with subscriptions, significant at the 0.001 level and rejecting the null hypothesis. “Euribor3m” interest rate, as expected, has a significant positive relationship with subscriptions, as does a successful previous campaign outcome, showing the importance of high interest rates and effective marketing campaigns. Both relationships are also significant at the 0.001 level. Previous contact during last campaign showed a negative relationship, significant at the 0.05 level. Days contacted were largely insignificant in terms of relationships with the dependent, however contact on Mondays was less likely to lead to subscriptions, showing a significant negative relationship at the 0.05 level. Certain months of contact were also less likely to have positive marketing outcomes, with customers contacted in May, November, October, and September less likely to subscribe to a term deposit (significant at the 0.001 level) while contact in March, from this model, is more likely to yield a positive outcome, showing a statistically significant positive relationship with subscriptions. Based on the model, telephone communication is less likely to be successful for direct marketing than mobile phone contact, showing a significant negative relationship with the binomial dependent. Customers in Blue Collar/ Manual Labour work, and less significantly service industries, are less likely than others to subscribe to a term deposit, based on the data. Regarding the other variables in the model, we cannot be confident that any significant positive or negative relationships exist with the dependent variable. The odds ratios measure the association between an exposure and an outcome, in this case subscription. In other words, the odds of subscription occurring based on exposure to a particular variable, compared to the odds when this variable is absent, or, if it is a binary categorical variable for example, replaced by the opposing variable. In this model, the odds for subscription to term deposits are 3.54 more likely for every unit increase in Eribor3m interest rate. On the other hand, contact by telephone the odds for subscription are 0.547 less than if contacted by mobile phone. The most significant odds ratio in the model is for the previous outcome variable, with odds for subscription from successful campaigns 7.17 larger than in the absence of success.

3.5 Assumptions and Predictive Accuracy

Various assumption checks and predictive accuracy tests were carried out on the final mode, with resluts shown in Figure 7.

**Figure 7** : Assumption Checks and Predictive Accuracy Statistics









The pseudo-R squared calculations for Hosmer & Lemeshow/Nagelkerke are within healthy ranges for the models’ fit to the data. Tests for standardised residuals found no violations of this assumption, with only 370/32950 observations > 1.96, below 5%. There were 0 observations with a Cooks distance > 1, and therefore no violations and no cases that have an undue influence on the model. There appears to be some possible issues with multicollinearity in the model, with euribor3m, emp\_var\_rate, nr\_employed, and month returning a GVIF of greater than 10, which may be fine for prediction but could be having an undue influence on the inferential model. The predictive accuracy of the model is high, at 94%, however the Kappa Value is a little under the usual good range of above 0.3, indicating just a fair model for predicting the target. The stark difference between these values shows a potential imbalance in the data.

3.5 Findings in Relation to Research

The findings from the model seem to suggest interest rates have a key influence on subscriptions. A study from Hughes (2016) supports this finding, stating that lower interest rates seem to push customers away from term deposits in favour or more on demand, less long-term solutions. On the other hand, other studies are in contradiction to the results found from this model. An analysis by Colaianni, Magdangal and Mitchell (2016), suggests that as numbers employed for marketing purposes increased, subscriptions to term deposits also rose. This may point towards QueensBank’s recruitment and training not being at the required level to see returns, or other external factors at the time the data was collected, however more raw data and analysis could clarify this relationship over time.

# Conclusions (250)

* 1. Implications for Theory and Practice

There are some implications in both theory and practice for QueensBank from the insights gained from this analysis. Training and better recruitment should always be a focus for the company, as findings suggest, increasing staff levels did not improve subscriptions and in fact had the opposite effect. Furthermore, interest rates should be closely tracked and followed, and campaign intensity should be increased based on times with higher interest rates, to maximise marketing potential. March also seems to be a popular month for subscriptions, so temporary recruitment could be adopted to take advantage of this period. The company may also see benefit in targeting customers that have university/professional qualifications, as well as students and those in retirement, as analysis shows some level of significance in relation to subscriptions.

* 1. Limitations and Recommendations for future Research

This study is limited in the sense that it is a snapshot of a period at QueensBank. Consumer Confidence, for example, rarely fluctuates in this data and is in the minus figures, so there can be no insight gained when consumer confidence is higher. Data spanning a greater length of time over several marketing campaigns and years would provide better insight. There is also some data conflicting with both intuition and other research and analysis in the area. This may be contextual, but a greater alignment of research could be achieved through further analysis over longer time-periods, and periods of both macroeconomic stability and instability. More data on both recruitment and training, call feedback, would also provide better performance-related insight. A study determining why certain economic and social variables had large odds ratios would also be useful.

# References

Coliani, G., Magdangal, J., and Mitchell, M. (2016) “Factors determining Term Deposit Purchases.” *How a Bank can get peoples Money,* 3-16

Hughes, T. (2016) “Are deposits safe under negative Interest Rates?” *Moody’s Analytics Risk Perspectives: The Convergence of Risk, Finance and Accounting*, Vol.7, p 1-5

Ling, X. and Li, C., 1998. “Data Mining for Direct Marketing: Problems and Solutions”. *In Proceedings of the 4th KDD conference,* AAAI Press, 73–79.

Moro, S., Laureano, R., and Cortez, P. (2011) “ Using data mining for bank direct marketing: an application of the CRISP-DM methodology” Available at <https://www.semanticscholar.org/paper/Using-data-mining-for-bank-direct-marketing%3A-an-of-Moro-Laureano/a175aeb08734fd669beaffd3d185a424a6f03b84#references> (Accessed 5th Dec 2020)

Ou, C., Liu, C., Huang, J. and Zhong, N. 2003. “On Data Mining for Direct Marketing”. *In Proceedings of the 9th RSFDGrC conference*, 2639, 491–498.

Page, C. and Luding, Y., 2003. “Bank manager’s direct marketing dilemmas – customer’s attitudes and purchase intention”. *International Journal of Bank Marketing* 21, No.3, 147–163.

# Appendix – R code

#this is what the code does....

#get working directory(correct)

getwd()

setwd("C:/Users/joede/OneDrive/Desktop/Statistics Assignment 2/Assignment2")

#install useful packages

install.packages("readxl")

install.packages("tidyverse")

install.packages("ggplot2")

install.packages("psych")

#read in the data

library(readxl)

data<- read\_excel("bank\_train.xlsx")

#summarise the data

summary(data)

#overview of data structure

str(data)

#convert characters to factors

data$job<-as.factor(data$job)

data$marital\_status<-as.factor(data$marital\_status)

data$education<-as.factor(data$education)

data$housing<-as.factor(data$loan)

data$loan<-as.factor(data$loan)

data$contact<-as.factor(data$contact)

data$month<-as.factor(data$month)

data$day\_of\_week<-as.factor(data$day\_of\_week)

data$poutcome<-as.factor(data$poutcome)

data$subscribed<-as.factor(data$subscribed)

data$default <- as.factor(data$default)

#exploration to identify outliers

boxplot(data$age)

hist(data$campaign)

boxplot(data$previous)

library(psych)

describe(data$previous)

describe(data$pdays)

describe(data$cons\_conf\_idx)

boxplot(data$cons\_conf\_idx)

#deal with numerical outliers/errors

data$age[data$age>80]<- NA

data$campaign[data$campaign>10]<- NA

data$duration[data$duration>500]<- NA

data$previous[data$previous>1]<- NA

data$cons\_conf\_idx[data$cons\_conf\_idx>1]<- NA

#descriptive statistics of clean data

summary(data)

str(data)

install.packages("Hmisc")

library(Hmisc)

describe(data)

#data visualisations using ggplot2

library(ggplot2)

#barchart showing subscriptions by month

ggplot(data) +geom\_bar(aes(x=subscribed, fill=month))+

labs(title= "Subscriptions by Month", x="Subscribed", y= "Frequency")

##scatterplot(jitter)- subscriptions by age and education type

ggplot(data=data, mapping=aes(x=age, y=education))+

geom\_jitter(mapping=aes(colour=subscribed))+

labs(title="Subscriptions by Age and Education Type", x= "Age", y="Education")+

labs(colour="Subscribed")

##density estimate- subscriptions by interest rates

ggplot(data=data) +

geom\_density(mapping = aes(x=euribor3m, fill=subscribed))+

labs(x="Interest Rate (Euribor 3 mth rate)", y="Frequency", title=" Subscription by Interest Rates")+

labs(fill="Subscribed")+

scale\_fill\_brewer(palette="Paired")

##scatterplot(jitter)- subscriptions Marketing and Contact

ggplot(data=data, mapping=aes(x=poutcome, y=previous))+

geom\_jitter(mapping=aes(colour=subscribed))+

labs(title="Subscriptions by Previous Marketing Campaign Outcomes and Previous Contact", x= "Previus Campaign Outcome", y="Times Contacted")+

labs(colour="Subscribed")+

scale\_color\_brewer(palette="Set1")

##barchart

ggplot(data=data)+

geom\_bar(mapping = aes(fill=subscribed, x= job))+

labs(fill="Subscribed",title="Subscriptions by Job Title",x="Job Title",y="Frequency")+

theme(axis.text.x=element\_text(angle = 90, hjust=1,vjust=0.5))+

scale\_fill\_brewer(palette="Reds")

##boxplot

ggplot(data=data, mapping=aes( x=cons\_conf\_idx, y=, fill=day\_of\_week))+

geom\_boxplot(outlier.shape = NA)+

facet\_wrap(facets = ~subscribed, nrow = 3, ncol=5)+

labs(fill="Day of Week", title="Effects of Consumer Confidence and Day of Week on Subscriptions", x= "Consumer Confidence Index")+

scale\_fill\_brewer(palette="Set1")

#Calculate measures of association

#Correlation with Categories

chisq.test(data$subscribed, data$job, correct=FALSE)

chisq.test(data$subscribed, data$contact, correct=FALSE)

chisq.test(data$subscribed, data$job, correct=FALSE)

chisq.test(data$subscribed, data$marital\_status, correct=FALSE)

chisq.test(data$subscribed, data$education, correct=FALSE)

chisq.test(data$subscribed, data$default, correct=FALSE)

chisq.test(data$subscribed, data$month, correct=FALSE)

chisq.test(data$subscribed, data$day\_of\_week, correct=FALSE)

chisq.test(data$subscribed, data$poutcome, correct=FALSE)

#t test for numeric variables and binary dependent

t.test(data$duration~data$subscribed)

t.test(data$campaign~data$subscribed)

t.test(data$previous~data$subscribed)

t.test(data$emp\_var\_rate~data$subscribed)

t.test(data$cons\_price\_idx~data$subscribed)

t.test(data$cons\_conf\_idx~data$subscribed)

t.test(data$euribor3m~data$subscribed)

t.test(data$nr\_employed~data$subscribed)

#read in test data

test<- read\_excel("bank\_test.xlsx")

#preprocessing test data to match training

test$age[test$age>80]<- NA

test$campaign[test$campaign>10]<- NA

test$duration[test$duration>500]<- NA

test$previous[test$previous>1]<- NA

test$cons\_conf\_idx[test$cons\_conf\_idx>1]<- NA

test$job<-as.factor(test$job)

test$marital\_status<-as.factor(test$marital\_status)

test$education<-as.factor(test$education)

test$housing<-as.factor(test$loan)

test$loan<-as.factor(test$loan)

test$contact<-as.factor(test$contact)

test$month<-as.factor(test$month)

test$day\_of\_week<-as.factor(test$day\_of\_week)

test$poutcome<-as.factor(test$poutcome)

test$subscribed<-as.factor(test$subscribed)

test$default <- as.factor(test$default)

#install required packages

install.packages("dplyr")

install.packages("caret")

library(dplyr)

library(caret)

#prepare data for modelling

data <- data %>% mutate\_if(is.character,as.factor)

data <- na.omit(data)

test<- na.omit(test)

#Logistic regression models

model1 <-glm(subscribed ~ nr\_employed + cons\_conf\_idx +euribor3m, data=data, family="binomial")

model2 <- glm(subscribed ~ nr\_employed + cons\_conf\_idx + euribor3m + emp\_var\_rate + previous + campaign, data=data, family="binomial")

formula <- subscribed ~ nr\_employed + cons\_conf\_idx + euribor3m + emp\_var\_rate + poutcome + campaign + day\_of\_week + month + housing + contact + education +job

final\_model <- glm(formula, data=data, family="binomial")

exp(final\_model$coefficients)

#review models

summary(model1)

summary(model2)

summary(final\_model)

#check predictive accuracy

logisticPseudoR2s <- function(LogModel) {

dev <- LogModel$deviance

nullDev <- LogModel$null.deviance

modelN <- length(LogModel$fitted.values)

R.l <- 1 - dev / nullDev

R.cs <- 1- exp ( -(nullDev - dev) / modelN)

R.n <- R.cs / ( 1 - ( exp (-(nullDev / modelN))))

cat("Pseudo R^2 for logistic regression\n")

cat("Hosmer and Lemeshow R^2 ", round(R.l, 3), "\n")

cat("Cox and Snell R^2 ", round(R.cs, 3), "\n")

cat("Nagelkerke R^2 ", round(R.n, 3), "\n")

}

logisticPseudoR2s(final\_model)

#assumption checks

data$predictedProbabilities <-fitted(final\_model)

data$standardisedResiduals <- rstandard(final\_model)

sum(data$standardisedResiduals >1.96)

summary(data$standardisedResiduals)

data$cook <- cooks.distance(final\_model)

sum(data$cook >1)

install.packages("car")

library(car)

vif(final\_model)

data$nr\_empLogInt <- log(data$nr\_employed)\*data$nr\_employed

data$confLogInt <- log(data$cons\_conf\_idx)\*data$cons\_conf\_idx

data$eurLogInt <- log(data$euribor3m)\*data$euribor3m

data$emp\_varLogInt <- log(data$emp\_var\_rate)\*data$emp\_var\_rate

formula2<- subscribed ~ nr\_employed + cons\_conf\_idx + euribor3m + emp\_var\_rate + poutcome + campaign + day\_of\_week + month + housing + contact + education +job

data$nr\_empLogInt + data$confLogInt + data$eurLogInt + data$emp\_varLogInt

modelcheck<-glm(formula2, data=data, family=binomial)

summary(modelcheck)

#prediction

data$subscribed = factor(data$subscribed,levels=c("no","yes"))

predictions <- predict(final\_model, test, type = "response")

class\_pred<- ifelse(predictions > .5, "yes", "no")

class\_pred = factor(class\_pred,levels=c("no","yes"))

confusionMatrix(table(class\_pred, test$subscribed))