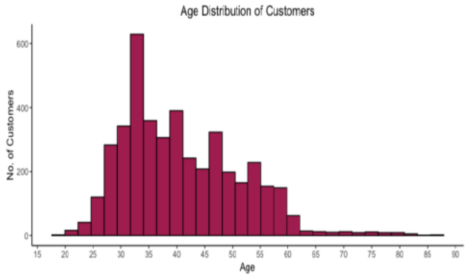
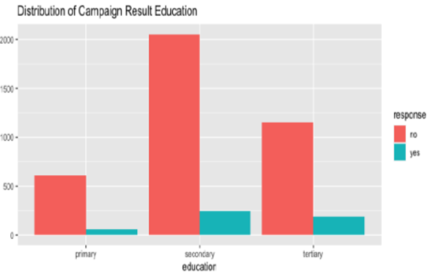
ST635 - Intermediate Statistical Modeling for Business
Fall 2023 - Group Project
Topic: Bank Marketing Campaign for Opening Term Deposit
Professor – Mengyan Li

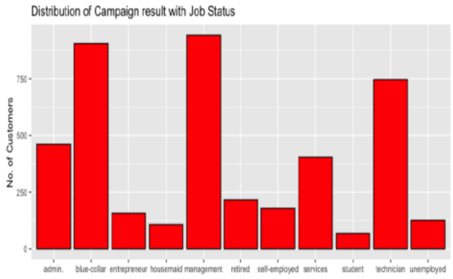
Mirjana Rajic
Adriana Choy
Shashank Sarangi
Sai Deekshitha

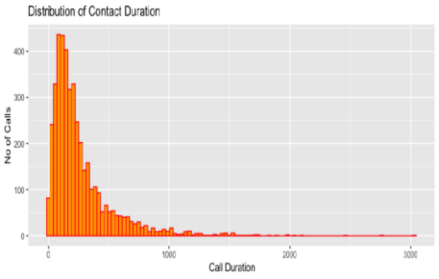
Exploratory Data Analysis

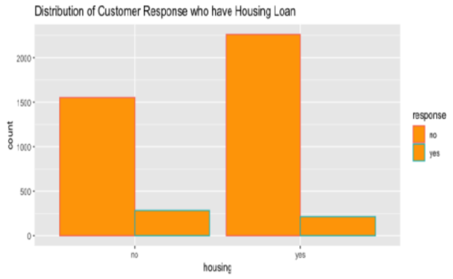
The dataset selected contains customer information such as age, marital status, job, education, loans, credit default, etc. This analysis focuses mainly on the marketing campaign efforts of the bank. The dataset includes 4521 rows of observations, 17 out of which 10 are categorical and 7 are numerical variables. The dataset has missing values: 38 in *job*, 187 in *education*, 1324 in *contact,* and 3705 in *poutcome*. Due to their small number, we removed the missing values from *jobs* and *education*. We then decided to perform a visualization analysis by plotting every variable in the dataset to understand the distribution. It was interesting to observe that most of the customers were between 25 and 60 years old, had blue-collar, admin, tech, and management jobs. More than 50% of customers possessed secondary education and used house loans, less than 15% used personal loans, and minimum credit defaulters. The EDA also showed that over 75% of the customers used cellular phones. When contacted by the bank, the communication lasted between 0 to 400 secs (6:40 mins), reflected by the right skewness in the plot, and more than 50% of the customer base is married, followed by single and then divorced.

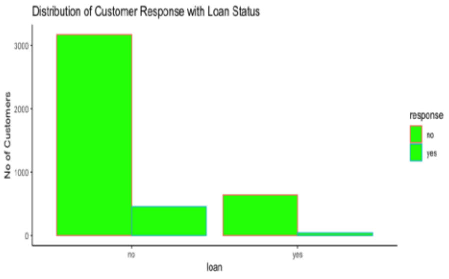












Poisson Regression

To understand the factors influencing the number of contacts during a marketing campaign, we decided to understand the Poisson distribution (a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space) by fitting a Poisson regression model and study if these events occur with a known constant mean rate and independently of the time since the last event. For example, '**month'**-seasonality may impact the campaign's effectiveness since certain times of the year may be better for outreach, '**contact'** - the method of contact (e.g., cellular or telephone) may influence the client's responsiveness to contact attempts, and '**pdays'** - the number of days since the last contact can indicate how recently the client has been exposed to the marketing messages: an extended gap merits more contacts.

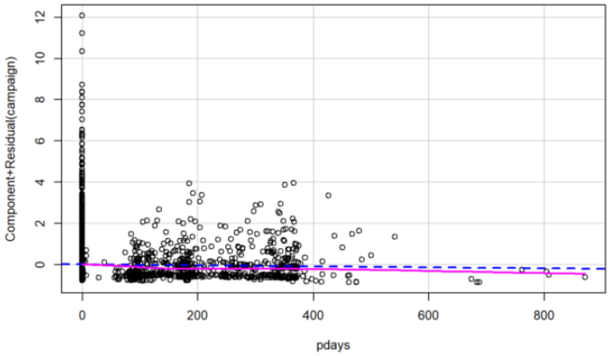
Similarly, '**duration'** - the number of seconds spent in the last contact with the client may influence fatigue or engagement; if the length of the previous contact is long, a contact may not be needed in the future. In summary, **month** accounts for seasonal effects, **contact** reports the usefulness of communication channels to reach the client, **duration** incorporates information on past responsiveness, and '**pdays**' captures recency and possible message decay, and previous handles potential contact fatigue. Together, they provide a broad view of the client's history and context to model the number of current contacts needed.

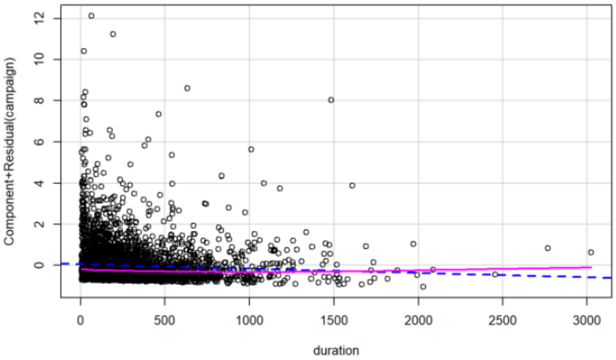
We then built a model to predict the number of marketing campaign successes based on the month of contact, days since previous contact, contact method, and duration of contact. The coefficient for the month suggests campaign success is significantly higher in August (coef = 0.705), July (0.630), June (0.412), March (0.297), May (0.191), and February (0.162) compared to the reference month of April. October (coef = -0.288) saw significantly fewer successes than April. The days since the previous contact coefficient (coef = -0.0004) indicate campaign success decreases the more extended the bank takes to contact the client again. Each additional day since the last contact reduces the log count of campaign contacts by 0.0004. Furthermore, the duration of the previous contact (coef = -0.0003) indicates that the length of prior contacts affects the responsiveness or need for more contacts during a campaign. Each additional second in the last contact with the client reduces the log count of campaign contacts by 0.0003. Lastly, the coefficient for contact suggests that contact count is significantly higher when done through the telephone (coef = 0.125) and other unknown channels (0.096) compared to cellular phone communication, which is the reference contact method in the model.

The model evaluation metrics indicate a good fit - the predictors improve on the null model, with lower residual deviance (8235.2 vs. 9275.1 for null) and adequate AIC (20253). The model identifies the month of contact, the length of the previous contact, and the method of contact and allows sufficient time to pass since the last contact as crucial optimization strategies for improving the campaign success rate. Contacting clients in the spring through summer months via telephone appears optimal when their previous contact was recent and relatively long in terms of time.

The model was further evaluated by a likelihood ratio test to compare the maximized likelihoods. With four degrees of freedom from the number of added predictors, the resulting p-value of 1.4e-222 indicates the Poisson model provides a substantially improved fit compared to the null. This suggests the included predictors contribute significant explanatory information for modeling campaign successes. The considerable, highly significant reduction in deviance provides evidence that incorporating factors like month, method of contact, and contact recency and duration improves our ability to model the number of successful marketing campaign contacts, compared to just using an intercept alone.

To assess the linearity assumptions of our model, we plotted component and residual plots where we could verify that numerical predictors such as pdays and duration have a linear relationship to the number of campaign contacts, as seen in the figure below.





The linearity assumption holds for pdays, but there is a slight non-linearity for the duration. We verified if this would be improved upon turning the duration variable into a polynomial, only to find that it adjusts best as is. In that sense, we concluded that both numerical predictors in our model have a linear relationship to the response variable, the number of campaign contacts, when holding all other predictors fixed.

We also verified the equidispersion in the model by calculating the ratio between deviance and degrees of freedom. The ratio in our model is 1.828, which indicates the distribution's mean is almost equal to the variance. This tells us that our model is well-specified and that its estimates are unbiased.

Conclusion

Based on our analysis, we recommend the bank change its customer targeting strategy through its future marketing campaigns by focusing more on customers between the ages of 25 and 50 years who are in management and administration jobs , predominantly during the month of august through winter via telephone appears optimal when their previous contact was recent and relatively long in terms of time. The customers should be segmented based on the account balances and be engaged through campaign calls not more than 3 to save time and effort.