Kyle Calabria

04/12/2018

Term Deposit Marketing Subscription Prediction

Milestone components

1. Define the problem
2. Identify your client
3. Describe your data set, and how you cleaned/wrangled it
4. List other potential data sets you could use
5. Explain your initial findings
6. Share the Capstone Project 2 code and milestone report related to Github repository

**Executive Summary**

**Problem**

The goal of this project is to predict whether a Portuguese banking institution’s clients will subscribe to their term deposit product. The data available for the predictions includes the customer demographic information, information related to the recent telemarketing campaign the bank has run, additional information pertaining to client contact and purchase history, and social and economic context data.

**Client**

The client is a Portuguese banking institution that would is representative of any financial services company that is attempting to sell a product to clients via a direct tele-marketing campaign.

**Data**

The data has 41,188 records and 21 features. The target variable is whether or not the customer subscribed to the loan which is a binary feature. The independent features are broken into four classes: customer demographic, marketing campaign results, other client, and social and economic context indicators. The customer demographic information includes age which is a numeric feature and the following six categorical features: marital status, education level, default status, housing loan and general loan status. The marketing campaign data includes the duration of the last phone call with the customer as a numeric feature as well as the following categorical features: type of phone reached (home or cell) and day of week and month of last contact. The other data related to the client includes whether or not the client subscribed to the term deposit marketed in the previous tele-marketing campaign by the bank as a categorical feature and the following numeric features: contacts during current campaign, days passed since last contact with client, and contacts performed to client before this campaign. Finally the following social and economic context indicators as numeric features are included: employment variation rate, consumer price index, consumer confidence index, Euribor three month rate, and number of employees.

**ETL Process**

The only feature with missing data was the feature called pdays which tracks the number of days since the client was last contacted from a previous campaign. 96.3% of the data were not contacted in a previous campaign which is recorded as the value 999. I handled this missing data by creating two variations of the data set. In one variation, given the high percentage of the data with missing values, I will trim the feature. In the second variation, I will keep the feature and leave the missing values as 999. Given that it is a numeric feature, If I imputed the value 0 instead, it would indicate to an algorithm that no time had passed since the last communication, whereas 999 is expressing the fact that no contact was made previously as a large integer which makes more sense.

All of the numeric data is in a good format in raw form for training models on. I hot-one encoded the following categorical features: job category, marital status, education level, loan default status, housing loan status, generic loan status, phone type, day of the week of last contact, outcome of previous marketing campaign, and the target variable tracking whether the client subscribed to the term deposit.

\*\*\* Default, housing, loan, - ‘yes’, ‘no’, ‘unknown’ \*\*\*

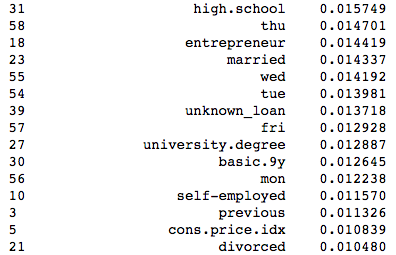
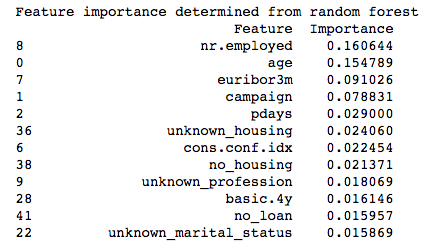
\*\*\* Poutcome - ‘failure’, ‘nonexistent’, ‘success’ \*\*\*

**Additional Data**

Timestamp data of the campaign related information would provide more detail for conducting time series analysis. A full voice recording of the phone call would provide a lot of data that a data scientist could use for analysis including natural language processing and sentiment analysis as well as other types of inflection and tone analysis. Geolocation or address of customers might be useful, although I would assume all of the clients would be in relatively close proximity to the bank. Geolocation still might be useful as different neighborhoods likely have different banking needs. Knowing other financial information related to the client could be highly significant indicators including data like other financial products purchased by the client, current assets held by the client at the bank, tax returns, transaction history, salary, current household income, etc. Other data pertaining to the client that may be interesting could include the length of time the client has been a client, social media data, size of the client’s family and whether they have a family member working at the bank.

**Exploratory Data Analysis**

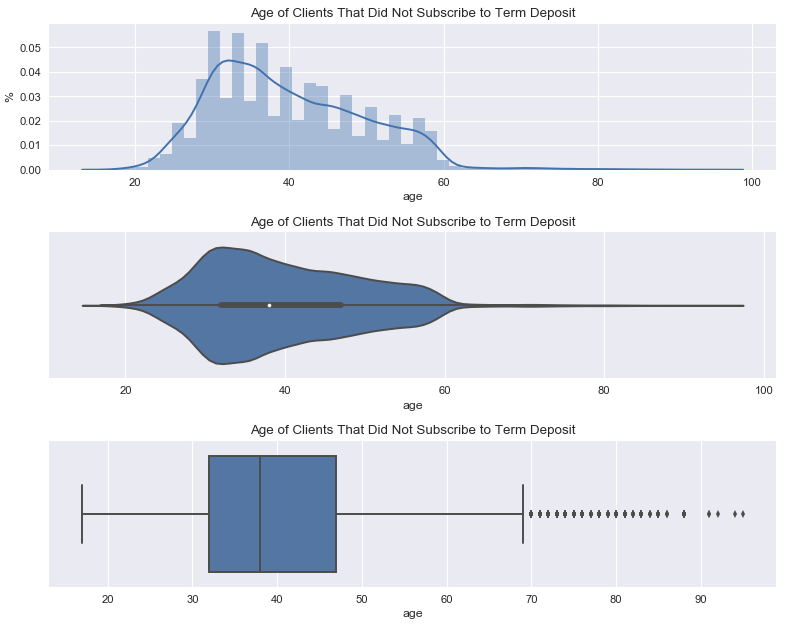
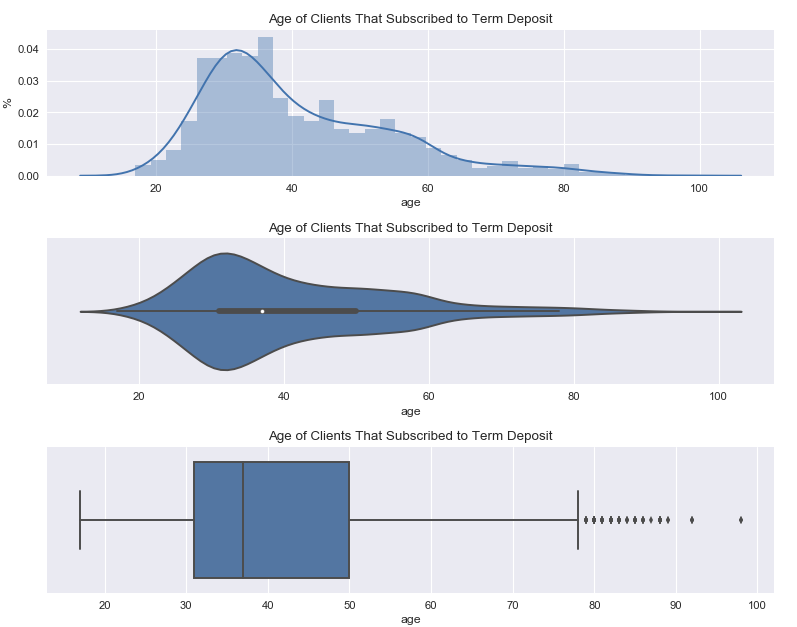
To orient myself of some of the important features to look at before beginning visual analysis of each individual feature, I ran a random forest model and created a feature importance data frame. Shown below are all of the features with a importance coefficient above .01. Social and economic context indicators played a more significant role in determining whether a client subscribed to the term deposit than I originally expected. My first assumption is that the term deposit rates are set based off of either one of the economic indicators here or another economic indicator which is highly correlated to the ones given here, likely LIBOR.

****

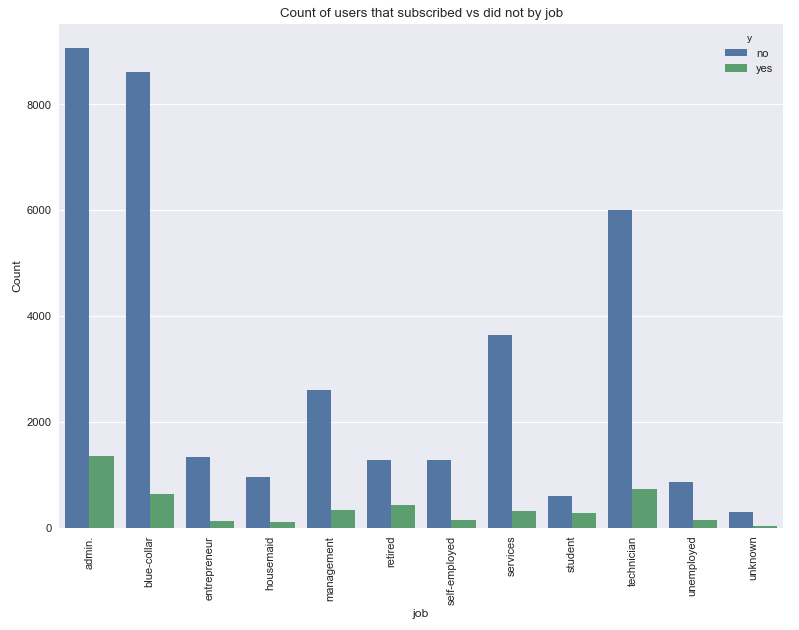
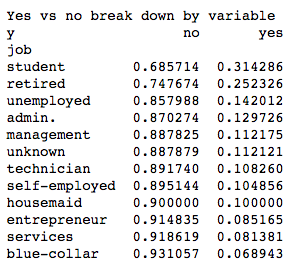
Age is the second most important feature which and the highest of all customer demographic information. This might be due to the fact that it is the only numeric feature of all customer demographic information thus all of the feature’s informational value is contained in a single column, whereas a feature like profession has had its informational value divided out amongst each individual profession column when it was hot-one encoded. Of the other demographic features, information regarding current lending habits, specifically mortgages is highly important. Another surprise is how many unknowns made the top twelve most important features including: unknown mortgage status, unknown professional status, and unknown marital status.

Of the campaign features, the number of contacts during the campaign was the most important followed directly by the number of previous days since the client was contacted in a previous campaign. I assume there is an optimal range of number of contacts that it would take to persuade a client to the term deposit, but after a threshold additional contacts would cause a decreased likelihood of subscription. For previous days feature, I expect that the fewer number of days since the last contact the better. I expected the results from the previous tele-marketing term deposit campaign i.e. whether or not the client subscribed to a term deposit during the previous tele-marketing campaign to be the single most important feature, but it is in fact not even in the top twenty most important features.

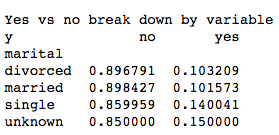
The above analysis and results are limited by the fact that they are specifically relevant to the model and don’t necessarily apply to other models or at an absolute level. Having said that, the feature importance is still likely to give me a rough idea of key features to explore further.



The counts by age for clients that subscribed vs those that did not follow different distributions. Clients that did subscribe follow a smoother distribution whereas clients that did not has a much more jagged distribution. The count of those that did not subscribe drops off sharply at 60 years of age whereas those that did subscribe slowly tapers down all the way past 80 years of age. Clients that did subscribe have an average age of 40.91, a standard deviation of 13.84 years and a median age of 37.00. Clients that did not subscribe have a mean of 39.9 years, a standard deviation of 9.90 years and a median of 38 years.

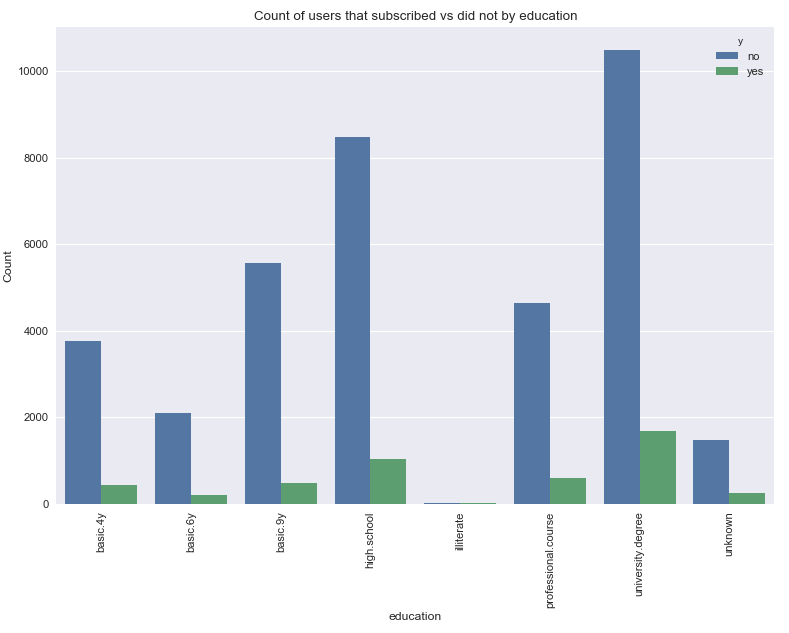


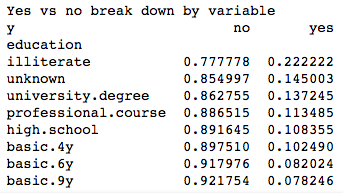
Students, retirees and unemployed have the highest proportion of subscriptions amongst professions, but only makeup a small percentage of the population. The fourth highest proportion of subscriptions amongst professions is admin., but this profession contains the largest percentage of the population. The profession with the lowest percentage of subscriptions is blue-collar, which is the second largest percentage of the population.

**Marital Status**

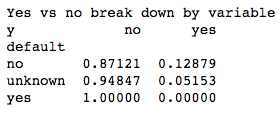
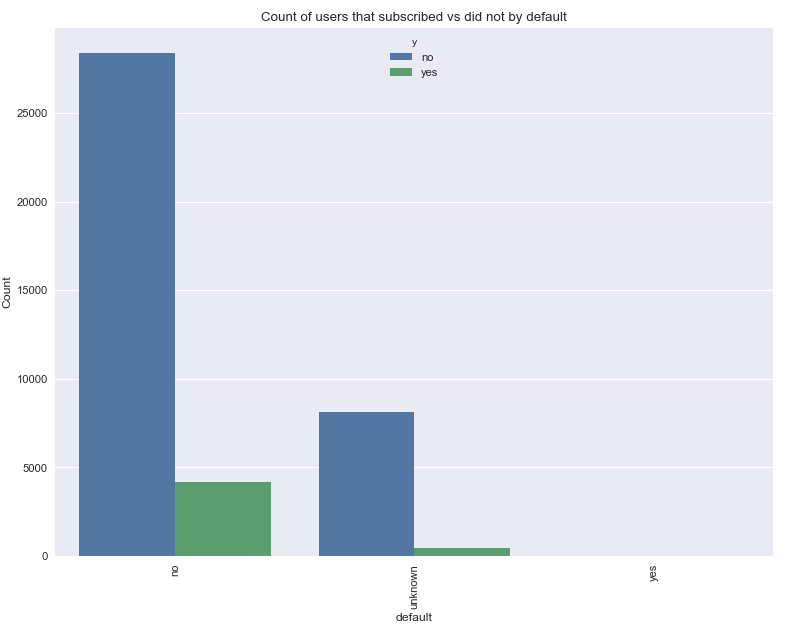
For material status there appears to be two groupings by proportion of clients to subscribe to the term deposit: those that have been married i.e. married and divorced and those that are either single or unknown. This pattern might indicate higher costs associated with raising a family resulting in the inability to invest money for the long term. On the other hand, one might expect a term deposit to be a good way to start a college savings fund for children.

**Education**

****

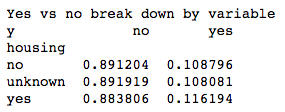
****

The most interesting insight for subscription result by education level is that those who are illiterate invest in term-deposits at a rate of nearly double all other education levels. This is followed by unknown, and then the two highest levels of education: university and professional. I am assuming that the illiterate category is suffering from sampling bias and if a larger sample of illiterate people were represented then the proportion that subscribed to the term deposit would be lower.

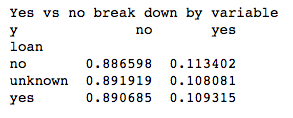
**Default Status**

Zero clients that are in default were investing in the term deposit which makes it a highly predictive variable; however, clients in default only makes up a negligible percentage of the population. There is also a significant margin between the proportions of clients known to not be in default vs clients whose default status is unknown.

**Mortgage Status**

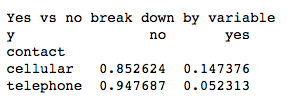
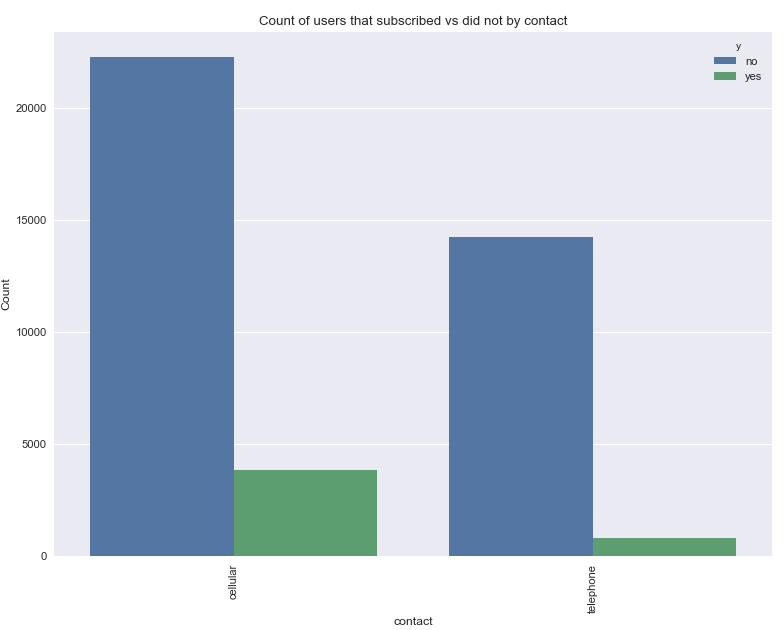
****

Whether or not the client had a mortgage had very little correlation on the proportion of clients within the category to subscribe to the term deposit.

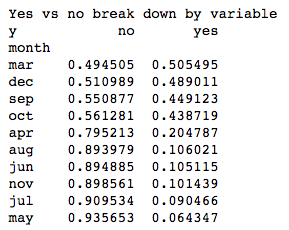
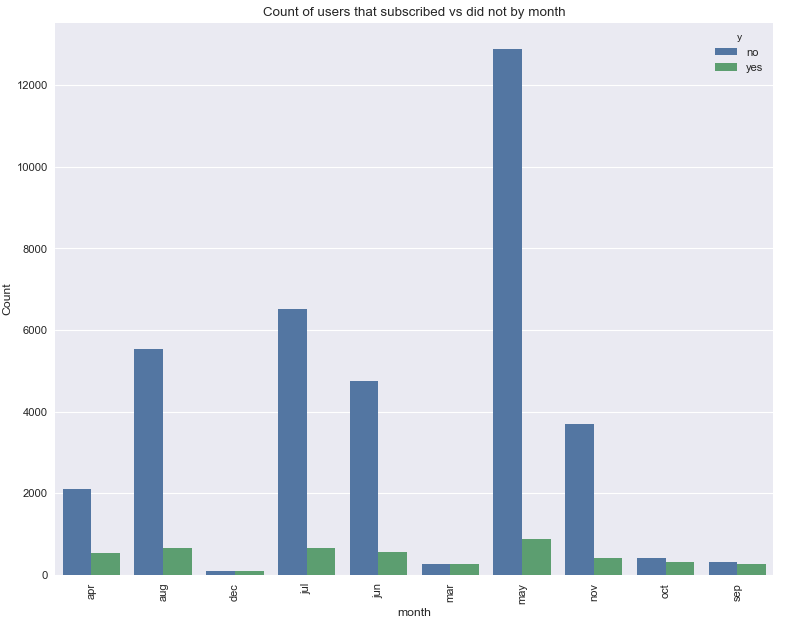
****

**Loan Status**

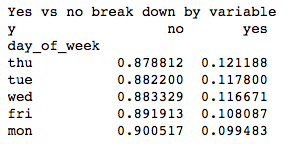
Loan status also had little correlation on the proportion of clients within the category to subscribe to the term deposit.

**Phone Type**

Roughly 63% of phone contact was made by cellular device as oppose to landline. If contact was made via cellular phone, they were nearly three times more likely to subscribe to the term deposit than if contact was made by landline.

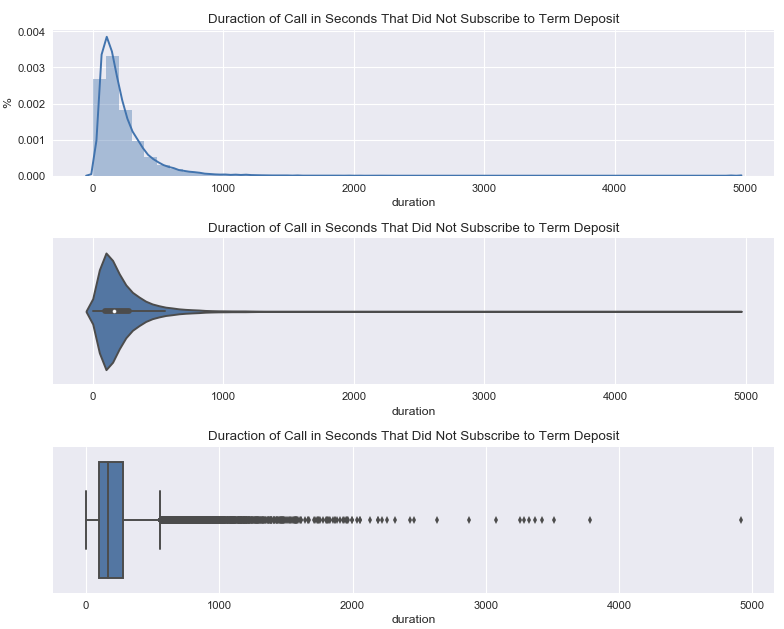
**Month**

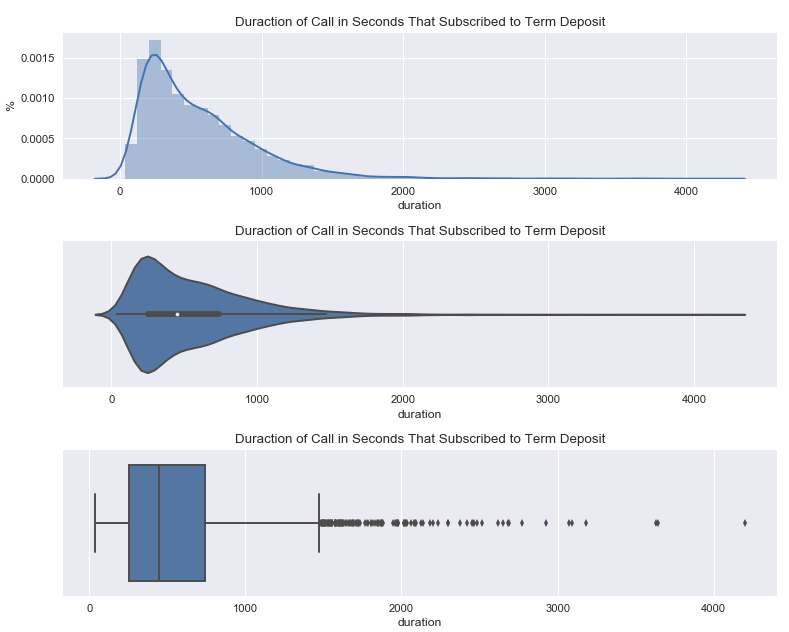
The month during the call was highly correlated with whether the client subscribed or not. March, December, September and October all had proportions of subscription above .4 whereas all other months were under .21. Given these proportions alone, someone might assume that the proportion of subscription of the overall population was much higher, but when looking closer, it is apparent that the months with the lowest proportion of subscriptions also make up the majority of the population with the bottom five months by proportion also being the top five highest counts of results.

****

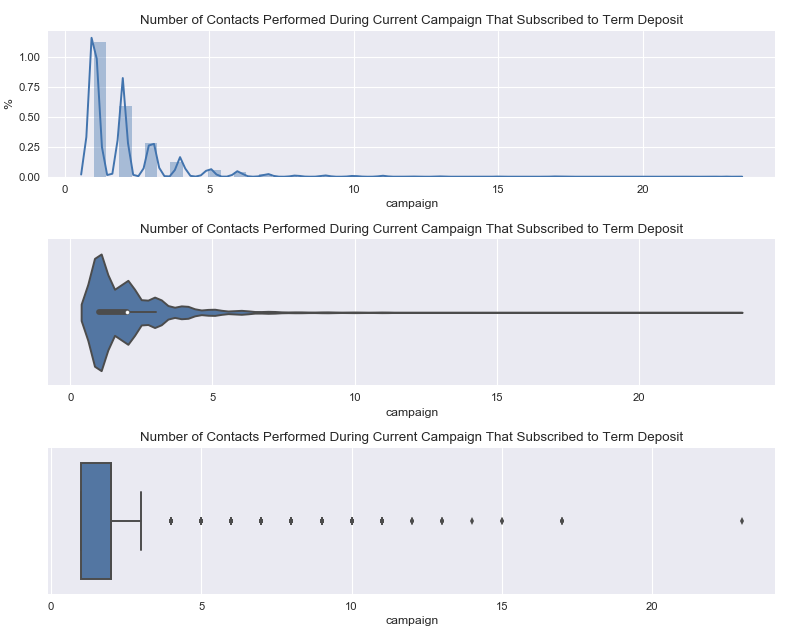
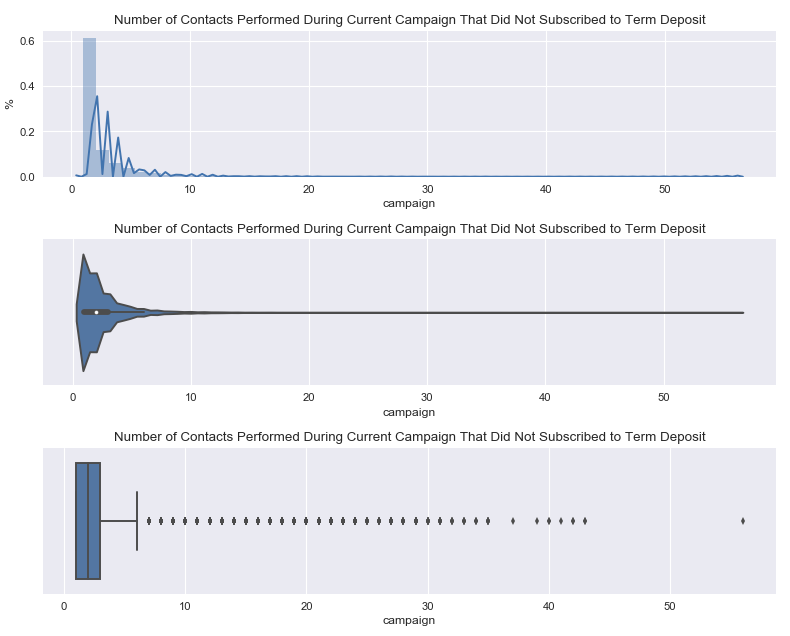
**Day of Week**

The day of the week that final contact to the client was made seems to have very little effect on whether or not they client subscribed to the term deposit.

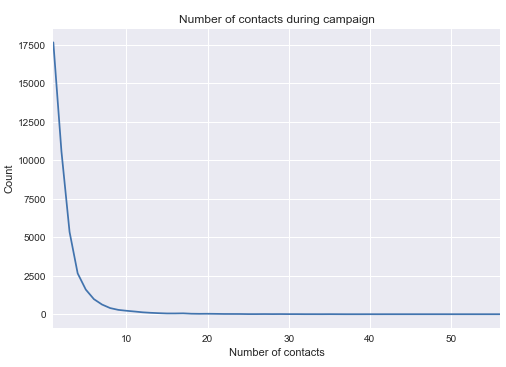
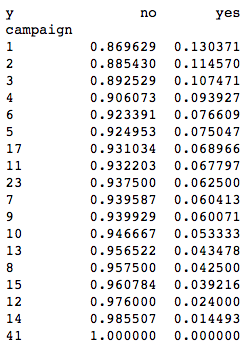
**Duration of Final Phone Call**

****

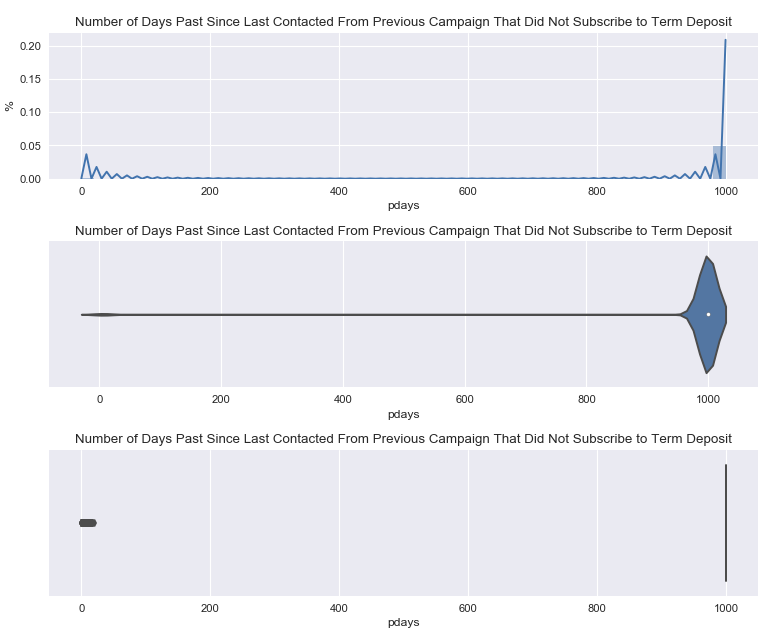
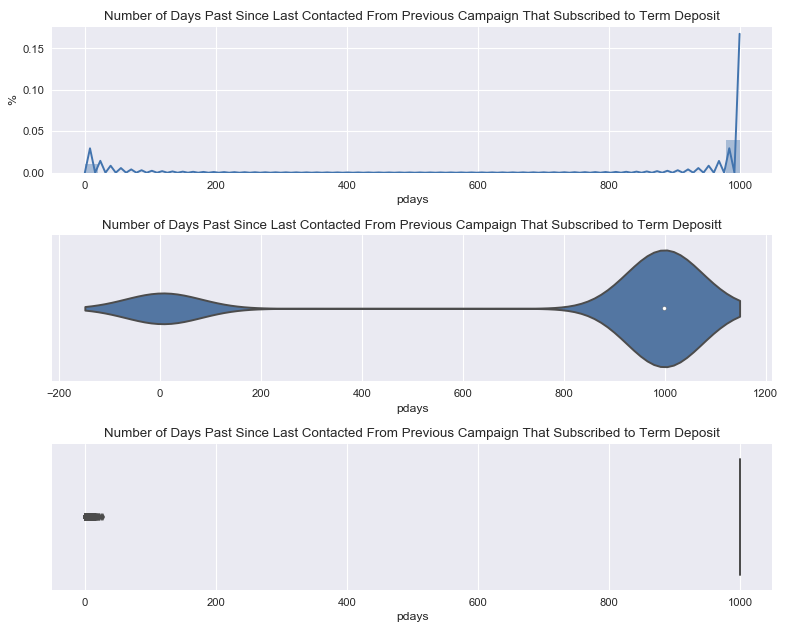
Unfortunately the duration feature cannot be used in model building, because this data is obtained at the same time as the client response to the target feature. Having said that, from the visualizations above, it is clear that longer phone calls are positively correlated with clients subscribing to the term deposit.

**Number of Contacts During Campaign**

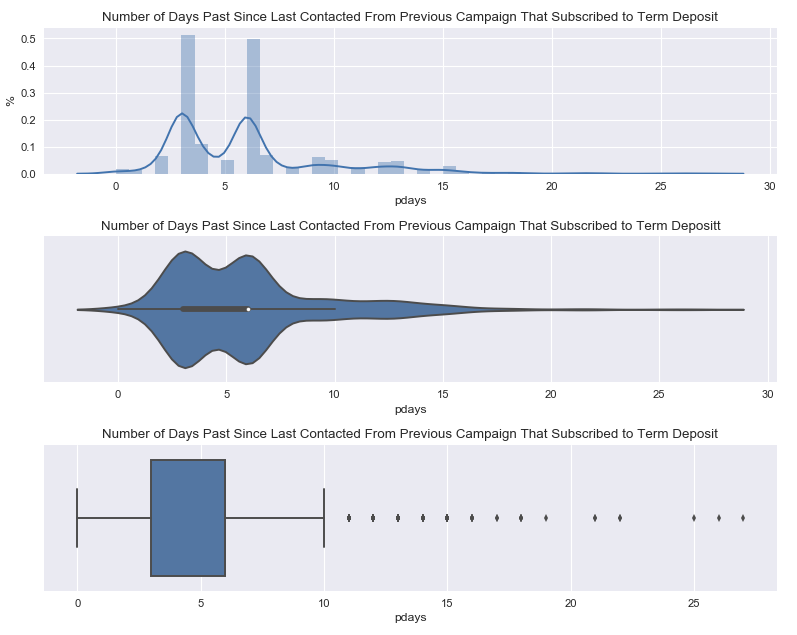
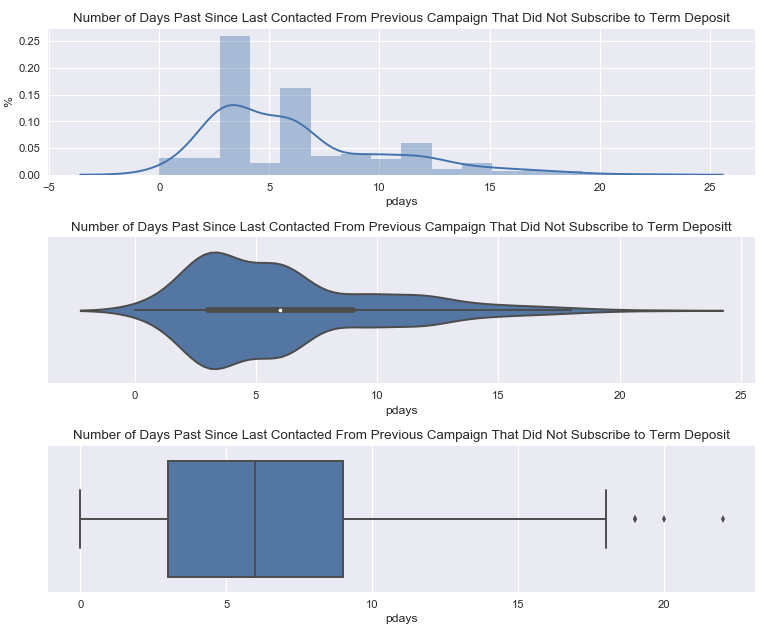
From the above visualizations of number of contacts during the campaign, it is evident that the fewer number of contacts made under four contacts, the more likely a client is to subscribe to a term deposit. Six contacts is just slightly more likely to produce a subscription than five contacts, and after six contacts there is no obvious pattern in number of contacts to whether a client will subscribe. Out of all the clients, 43% were contacted once, 25% were contacted twice, and 13% were contacted three times and 91% of clients were contacted five times or less. Clients that were contacted once or twice are the only number of contacts to have proportions higher than the total proportion of 11.27%.

****

**Number of days since client was last contact from previous campaign**

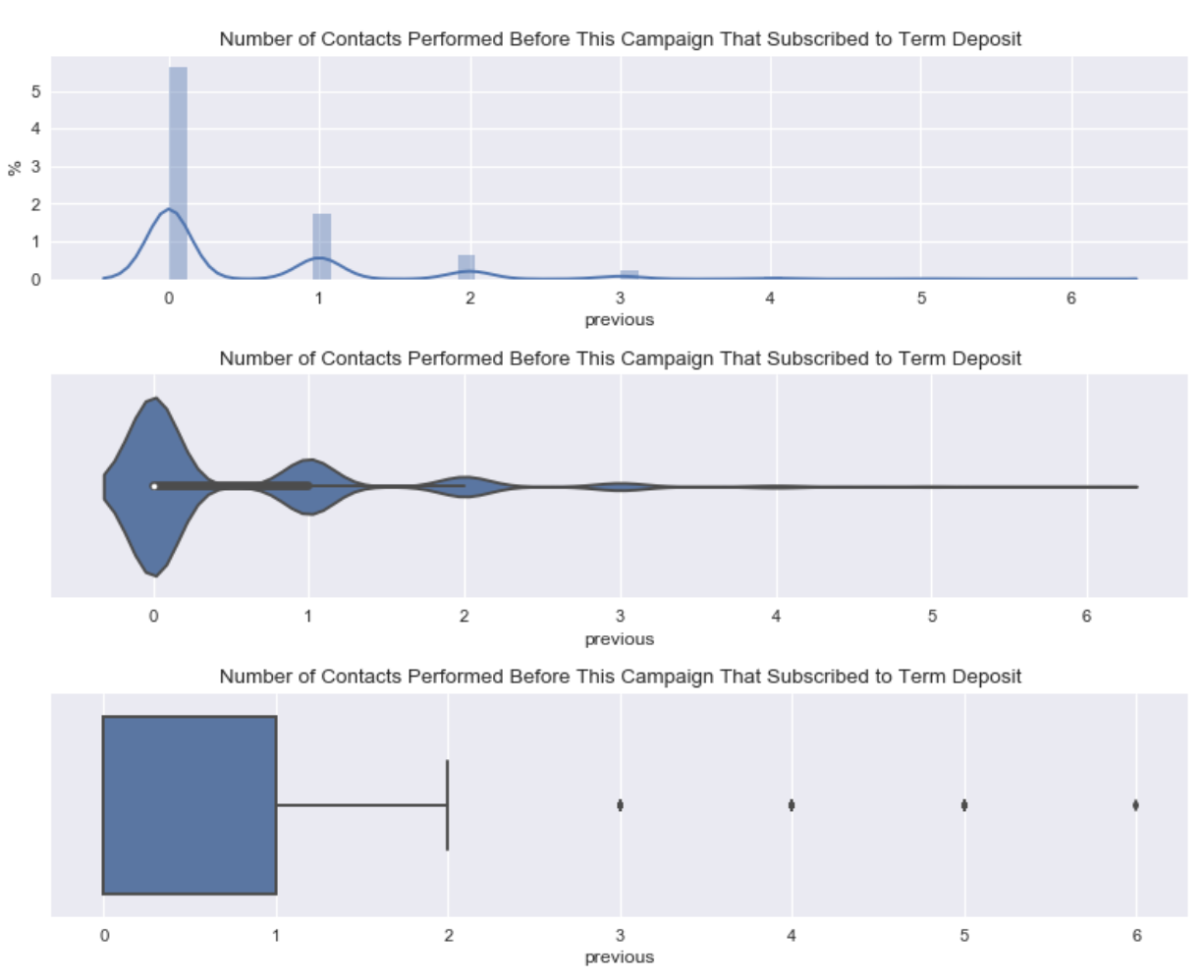
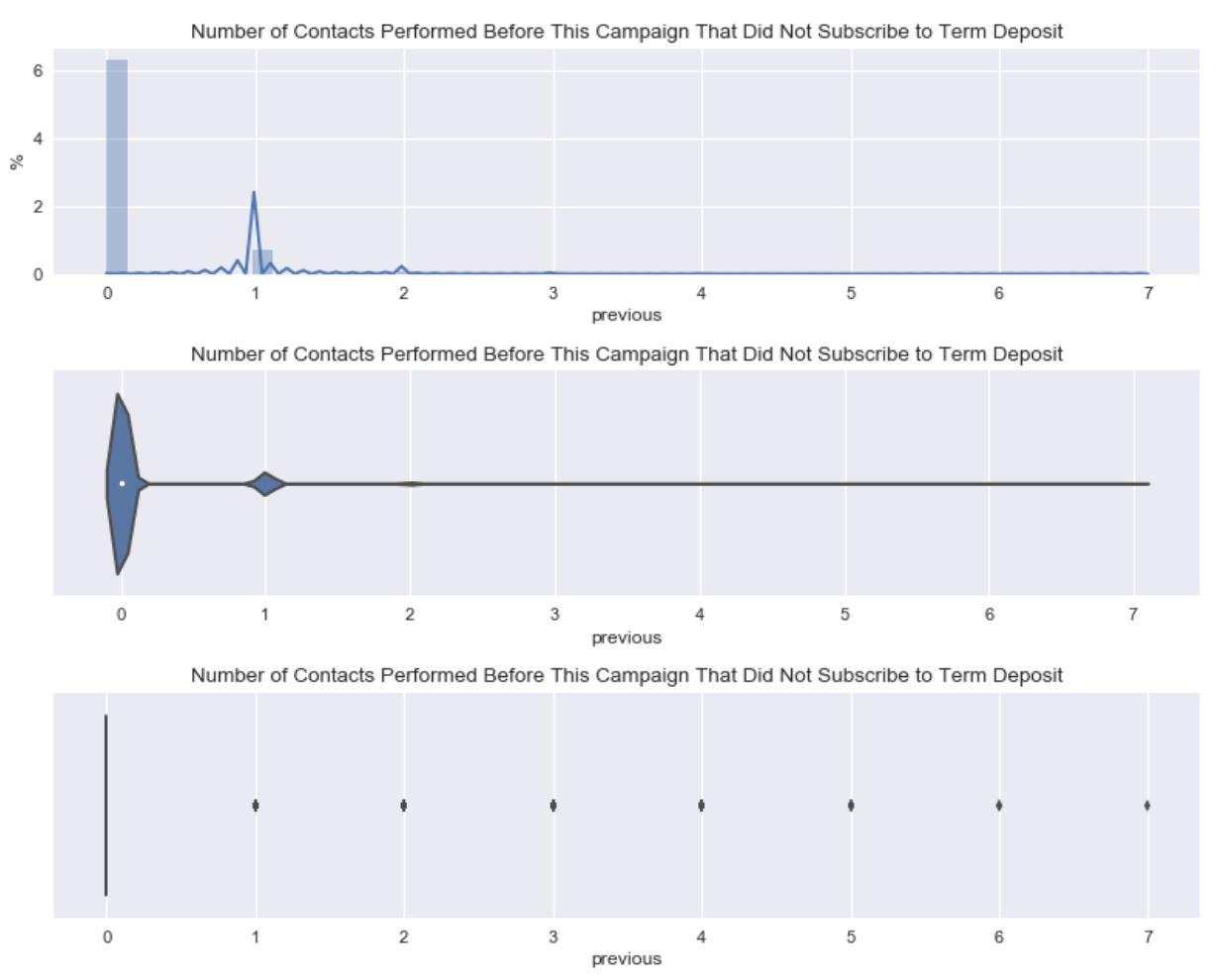
****

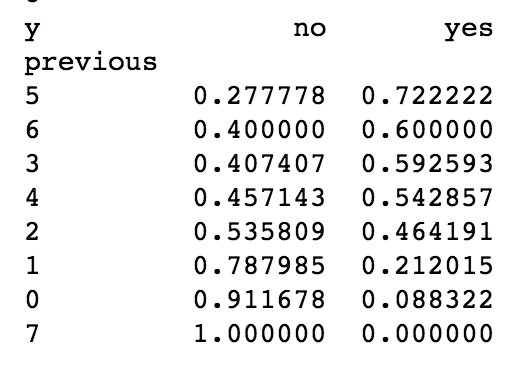
From the above plots it is evident that the distribution is lopsided due to all of the clients that have not previously been contacted which is encoded as 999. Below are the same plots with the clients that have not been contacted removed.

****

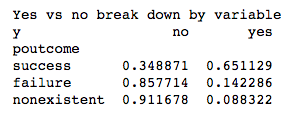
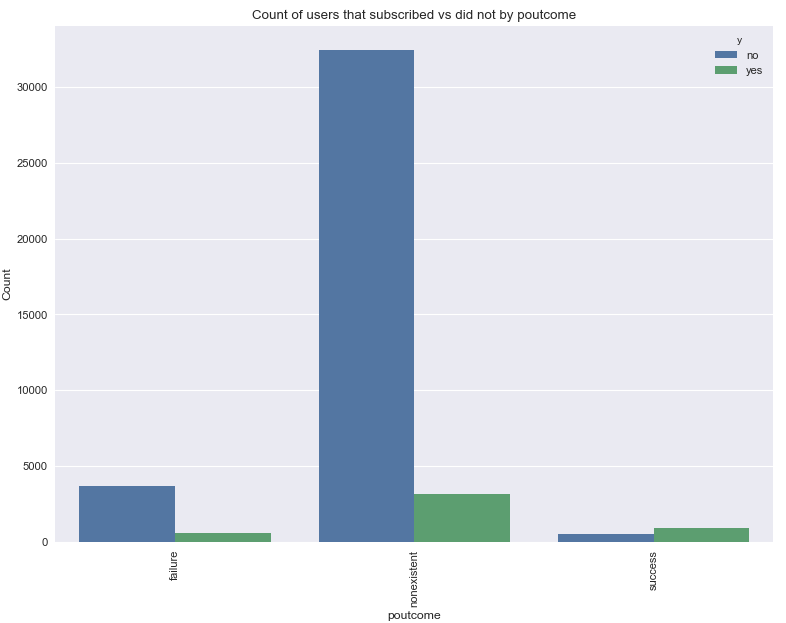
Due to the fact that 96% of the data contained values of 999 representing no previous contact, the visualizations directly above provide little insight for the entire dataset.

**Number of Contacts Performed Before this Campaign**

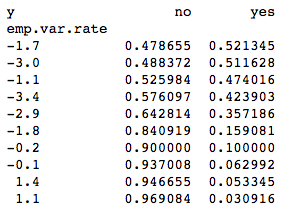
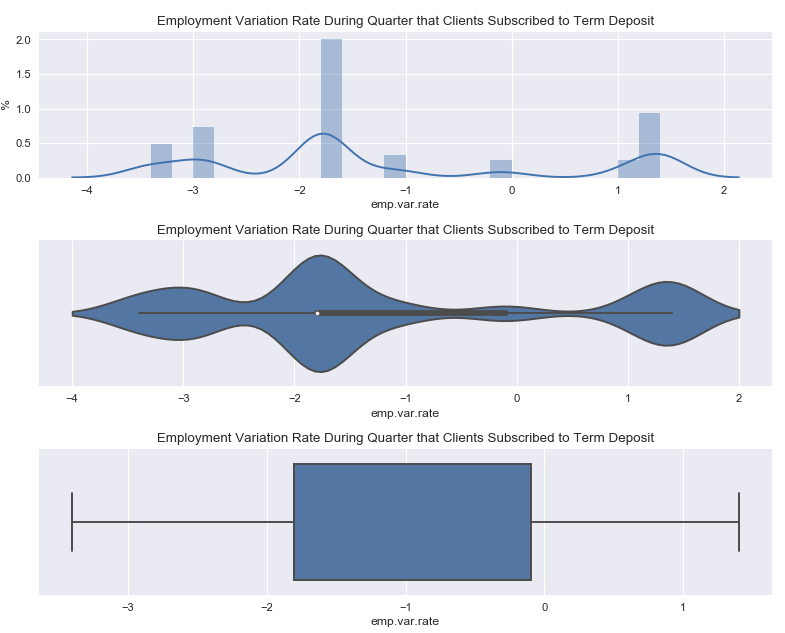
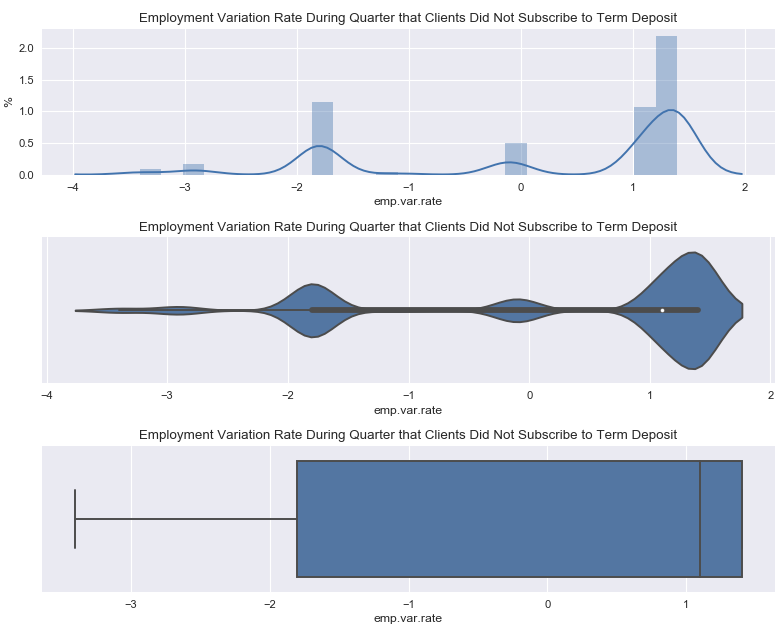


****

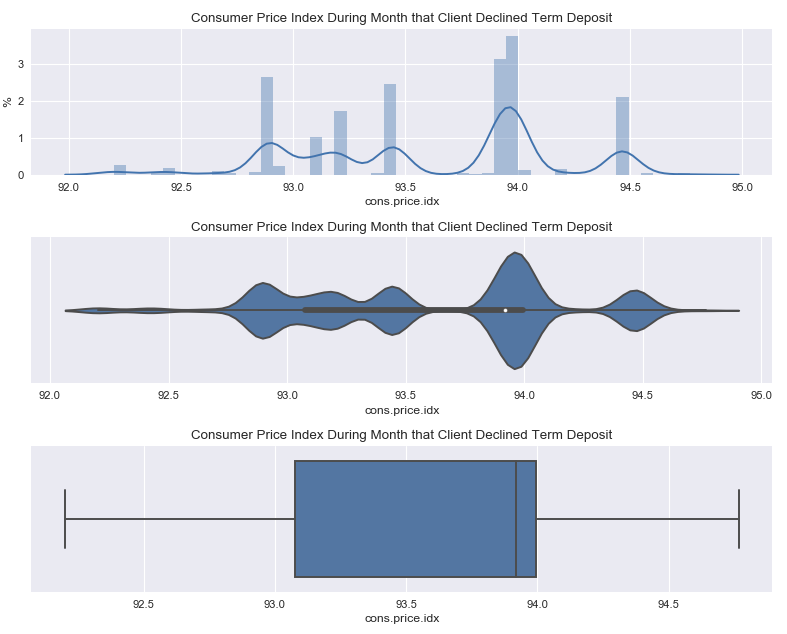
**Outcome of Previous Tele-marketing Campaign**

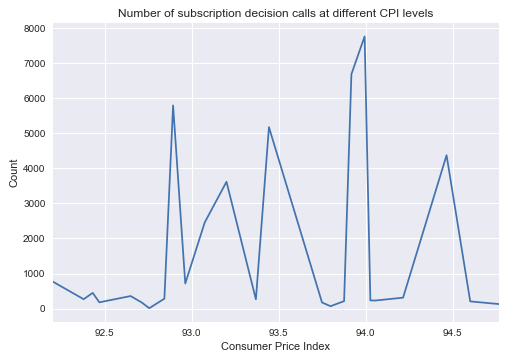
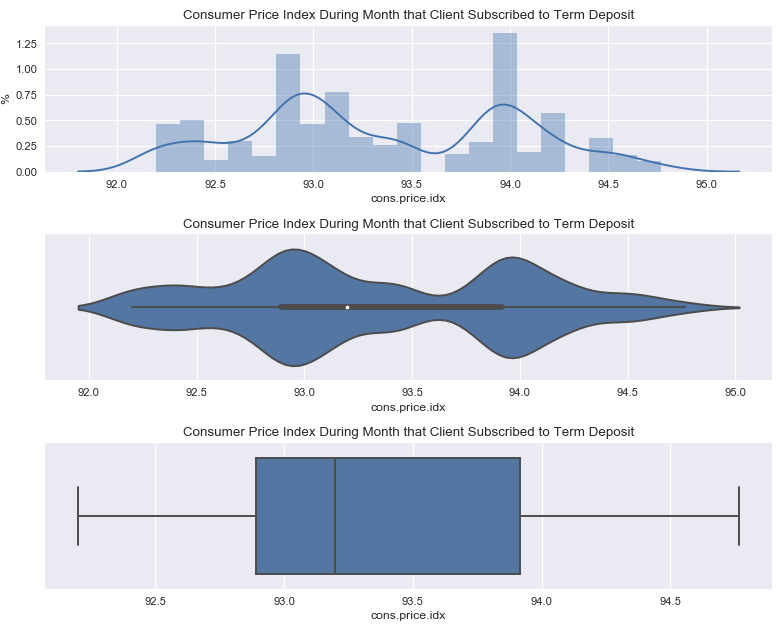
****

The proportion of term deposit subscriptions for clients having subscribed to a term deposit in a previous marketing campaign is four times higher than those that did not sign up in a previous campaign and over six times higher than had the client not been marketed a term deposit in the past at all.

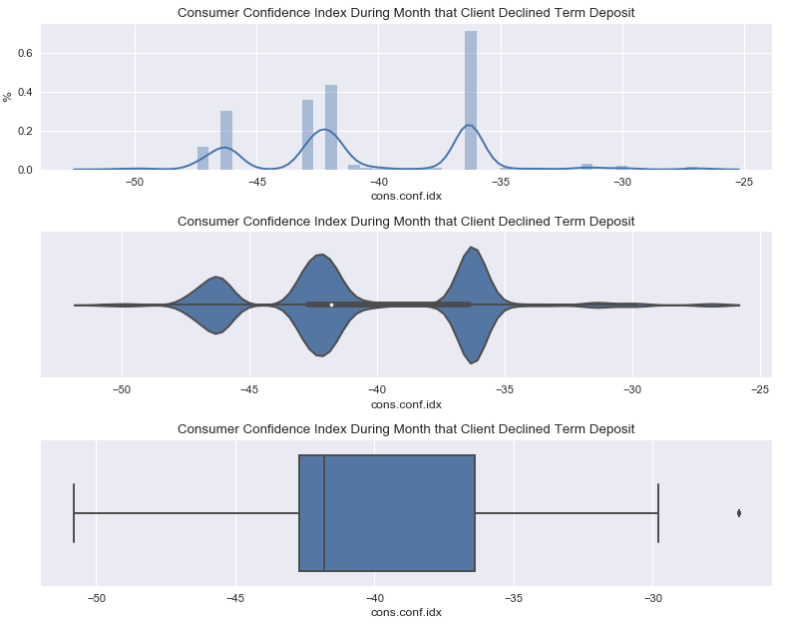
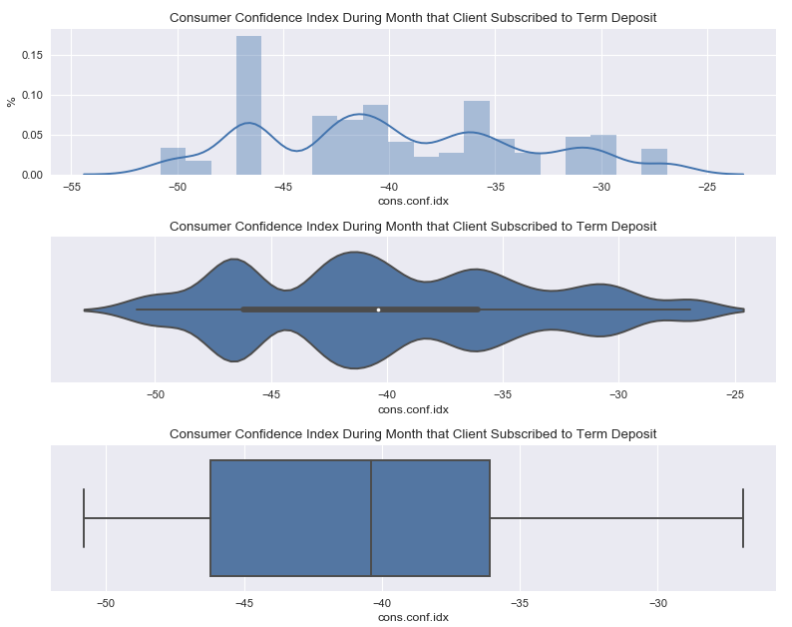
**Employee Variation Rate**

The general trend visible from the employee variation rate visualizations and table is that as employee variation increases, clients are less likely to be investing money in term deposits.

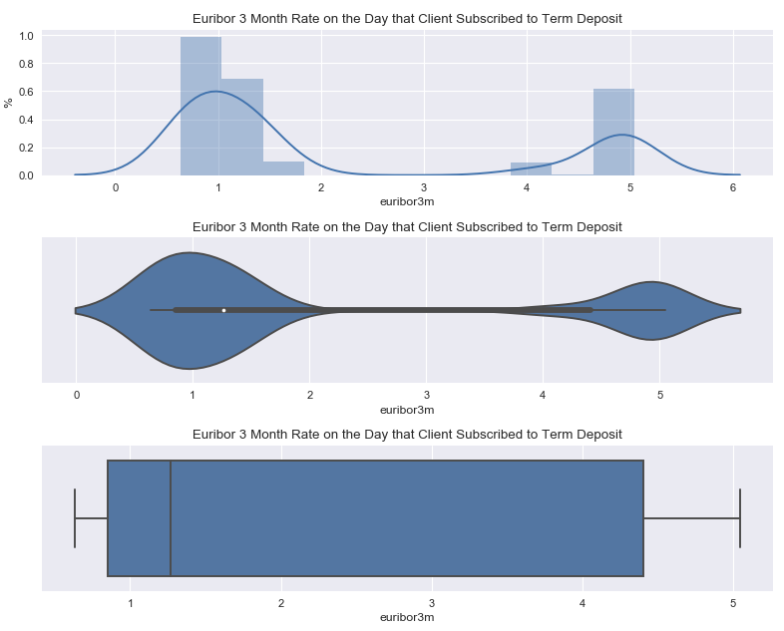
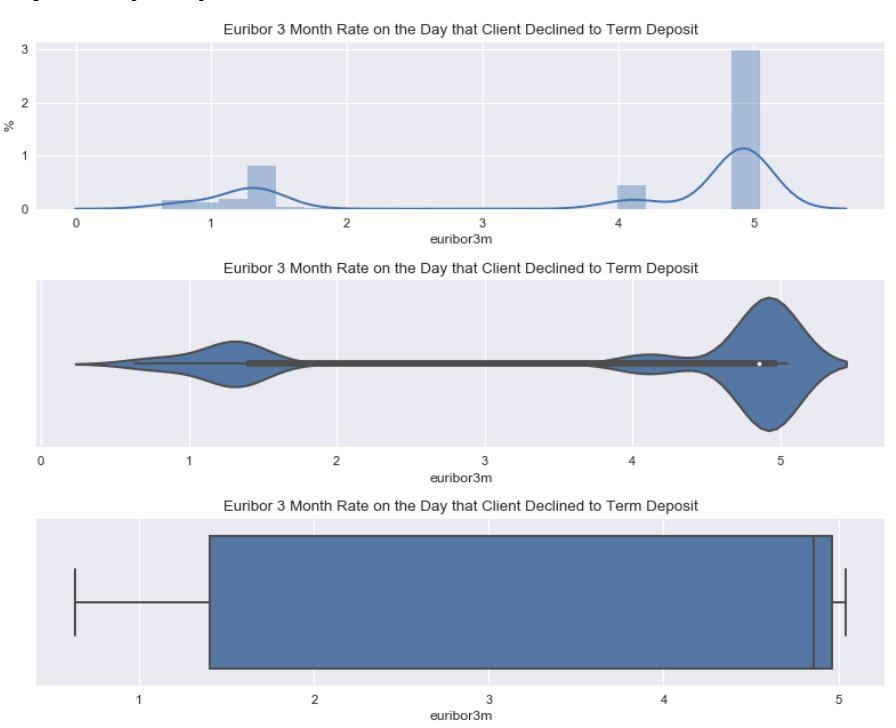
**Consumer Price Index**

****

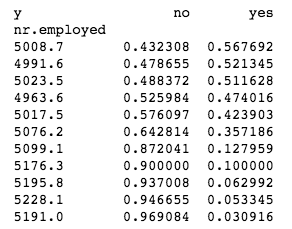
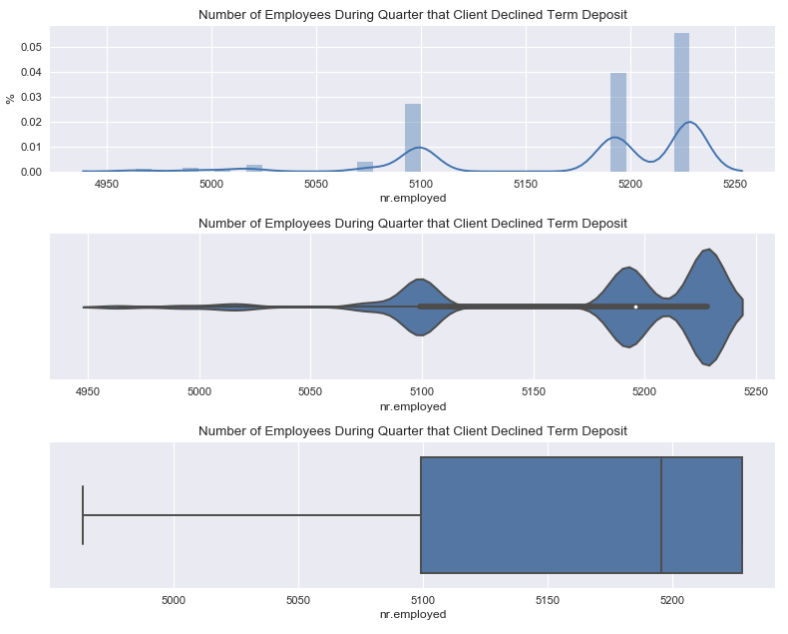
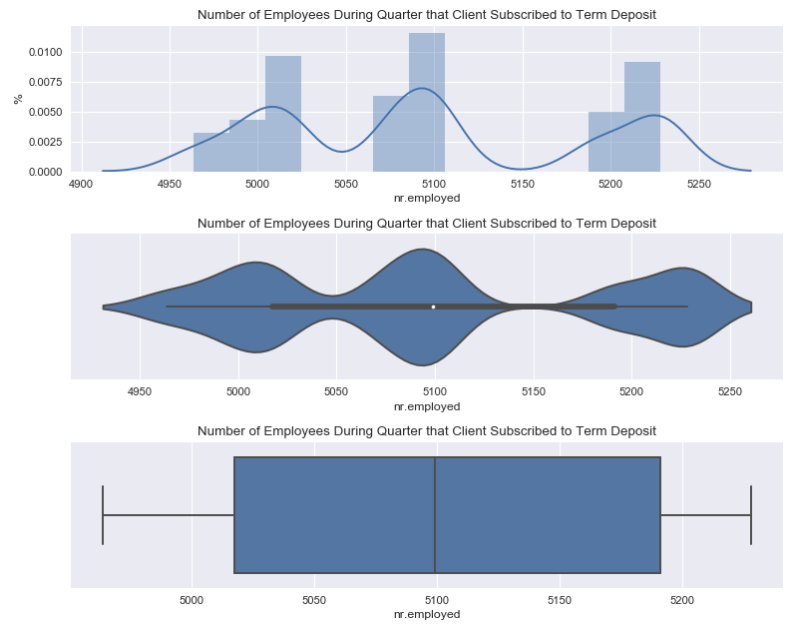
The distribution of clients that subscribed given different levels of consumer price index is distributed bi-modally with a slightly larger grouping of subscriptions at the 93.0 mode. Overall the average consumer price index is lower for those that subscribed than for those who did not. The distribution of clients that declined to subscribe given different levels of consumer price index is distributed multi-modally with the largest mode being at 94.0. Both subscribed and declined clients had a mode at 94.0 which suggests that a large portion of final subscription decision calls were made when the consumer price index was at 94.0. This is backed up by the distribution of total calls at varying CPI levels.

**Consumer Confidence Index**

The distribution of clients that subscribed given different levels of consumer confidence index is distributed multi-modally with a an average just below -41. Overall the average consumer confidence index is lower for those that declined than for those who did. The distribution of clients that declined to subscribe given different levels of consumer price index is distributed tri-modally with three distinct modes at -47, -42.5 and -37. The distribution of declines is much less smooth than for those that subscribed. Out of 26 different CCIs the top five by count make up 86% of all declines, whereas for subscribed term deposits the top five only make up 47%.

**Euribor Three Month Rate**

It is evident that the Euribor three month rate for both clients that subscribed and that declined is bimodal with modes for both being near one and five. This suggests the majority of all final decision calls occurred at these two euribor rates. It is also evident that as euribor rate increases there are more declines and as it decreases there are more subscriptions.

**Number of Employees During Quarter**

Number of employees during the quarter appears to more strongly correlate with clients that declined the subscription i.e. more clients declined when the number of employees during a quarter was higher. The number of employees during the quarters when clients subscribed seems to be tri-modal and be less correlated with three similar counts at each mode.

**Model Building and Machine Learning Analysis**

Loss Function

Feature Engineering

Feature Selection

Methodology

Feature Importance

**Recommendations and Further Learning**

**Appendix**

aTarget variable (1) - ‘y’ : binary

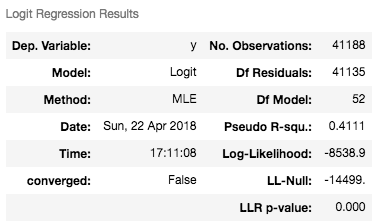
Customer demographic (7) ‘age’: numeric, ‘job’ : categorical, ‘marital’ : categorical, ‘education’ : categorical, ‘default’ : categorical, ‘housing’ : categorical, ‘loan’ : categorical

Marketing Campaign (4) ‘contact’ : categorical, ‘month’ : categorical, ‘day\_of\_week’ : categorical, ‘duraction’ : numeric,

\*\*\***Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.\*\*\***

Other (4) ‘campaign’ : numeric, ‘pdays’ : numeric, ‘previous’ : numeric, ‘poutcome’ : categorical

\*\*\*999 - client was not previously contacted for ‘pdays’\*\*\*

Social / Economic Context Arributes (5): ‘emp.var.rate’ : numeric, ‘cons.price.idx’ : numeric, ‘cons.conf.idx’ : numeric, ‘euribor3m’ : numeric, ‘nr.employed’ : numeric

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **coef** | **std err** | **z** | **P>|z|** | **[0.025** | **0.975]** |  |
| **age** | 0.0002 | 0.002 | 0.081 | 0.936 | -0.005 | 0.005 |
| **duration** | 0.0047 | 7.46E-05 | 63.108 | 0.000 | 0.005 | 0.005 |
| **campaign** | -0.0401 | 0.012 | -3.473 | 0.001 | -0.063 | -0.017 |
| **pdays** | -0.0009 | 0.000 | -4.326 | 0.000 | -0.001 | -0.001 |
| **previous** | -0.0628 | 0.059 | -1.062 | 0.288 | -0.179 | 0.053 |
| **emp.var.rate** | -1.7576 | 0.142 | -12.380 | 0.000 | -2.036 | -1.479 |
| **cons.price.idx** | 2.1905 | 0.252 | 8.679 | 0.000 | 1.696 | 2.685 |
| **cons.conf.idx** | 0.0207 | 0.008 | 2.664 | 0.008 | 0.005 | 0.036 |
| **euribor3m** | 0.3316 | 0.130 | 2.551 | 0.011 | 0.077 | 0.586 |
| **nr.employed** | 0.0054 | 0.003 | 1.738 | 0.082 | -0.001 | 0.012 |
| **entrepreneur** | -7.7378 | 8.86E+07 | -8.73E-08 | 1.000 | -1.74E+08 | 1.74E+08 |
| **blue-collar** | -7.9726 | 6.48E+07 | -1.23E-07 | 1.000 | -1.27E+08 | 1.27E+08 |
| **services** | -7.9159 | 7.33E+07 | -1.08E-07 | 1.000 | -1.44E+08 | 1.44E+08 |
| **management** | -7.7621 | nan | nan | nan | nan | nan |
| **unemployed** | -7.7940 | 6.44E+07 | -1.21E-07 | 1.000 | -1.26E+08 | 1.26E+08 |
| **student** | -7.4520 | 9.29E+07 | -8.02E-08 | 1.000 | -1.82E+08 | 1.82E+08 |
| **self-employed** | -7.8956 | nan | nan | nan | nan | nan |
| **housemaid** | -7.8777 | 7.69E+07 | -1.02E-07 | 1.000 | -1.51E+08 | 1.51E+08 |
| **admin.** | -7.5345 | 9.73E+07 | -7.75E-08 | 1.000 | -1.91E+08 | 1.91E+08 |
| **technician** | -7.7518 | nan | nan | nan | nan | nan |
| **retired** | -7.7167 | 7.09E+07 | -1.09E-07 | 1.000 | -1.39E+08 | 1.39E+08 |
| **unknown\_profession** | -7.8083 | 7.1E+07 | -1.1E-07 | 1.000 | -1.39E+08 | 1.39E+08 |
| **divorced** | -23.3187 | 2.28E+06 | -1.02E-05 | 1.000 | -4.46E+06 | 4.46E+06 |
| **coef** | **std err** | **z** | **P>|z|** | **[0.025** | **0.975]** |  |
| **married** | -23.3214 | 1.66E+07 | -1.41E-06 | 1.000 | -3.25E+07 | 3.25E+07 |
| **single** | -23.2628 | 1.53E+07 | -1.52E-06 | 1.000 | -2.99E+07 | 2.99E+07 |
| **unknown\_marital\_status** | -23.2892 | 3.16E+06 | -7.37E-06 | 1.000 | -6.19E+06 | 6.19E+06 |
| **illiterate** | -11.8638 | 4.77E+07 | -2.49E-07 | 1.000 | -9.35E+07 | 9.35E+07 |
| **basic.6y** | -11.7412 | 5.64E+07 | -2.08E-07 | 1.000 | -1.11E+08 | 1.11E+08 |
| **university.degree** | -11.8648 | 3.23E+07 | -3.67E-07 | 1.000 | -6.33E+07 | 6.33E+07 |
| **basic.4y** | -11.8152 | 3.72E+07 | -3.17E-07 | 1.000 | -7.3E+07 | 7.3E+07 |
| **basic.9y** | -10.7972 | 8.6E+06 | -1.26E-06 | 1.000 | -1.69E+07 | 1.69E+07 |
| **high.school** | -11.7489 | 4.01E+07 | -2.93E-07 | 1.000 | -7.85E+07 | 7.85E+07 |
| **unknown\_education** | -11.6680 | 3.21E+07 | -3.63E-07 | 1.000 | -6.29E+07 | 6.29E+07 |
| **professional.course** | -11.7141 | 2.61E+07 | -4.49E-07 | 1.000 | -5.11E+07 | 5.11E+07 |
| **yes\_default** | -26.3821 | 3.69E+07 | -7.15E-07 | 1.000 | -7.23E+07 | 7.23E+07 |
| **unknown\_default** | -26.6824 | nan | nan | nan | nan | nan |
| **no\_default** | -40.1394 | 5.98E+07 | -6.71E-07 | 1.000 | -1.17E+08 | 1.17E+08 |
| **yes\_housing** | -31.0636 | 4.08E+07 | -7.62E-07 | 1.000 | -7.99E+07 | 7.99E+07 |
| **unknown\_housing** | -31.1004 | nan | nan | nan | nan | nan |
| **no\_housing** | -31.0684 | 3.81E+07 | -8.15E-07 | 1.000 | -7.47E+07 | 7.47E+07 |
| **yes\_loan** | -31.0361 | nan | nan | nan | nan | nan |
| **unknown\_loan** | -31.0927 | nan | nan | nan | nan | nan |
| **no\_loan** | -31.0877 | 5.1E+06 | -6.09E-06 | 1.000 | -1E+07 | 1E+07 |
| **telephone** | -45.5373 | 2.73E+07 | -1.67E-06 | 1.000 | -5.35E+07 | 5.35E+07 |
| **cellular** | -46.1833 | 2.65E+07 | -1.74E-06 | 1.000 | -5.19E+07 | 5.19E+07 |
| **coef** | **std err** | **z** | **P>|z|** | **[0.025** | **0.975]** |  |
| **oct** | -9.6193 | 5.37E+07 | -1.79E-07 | 1.000 | -1.05E+08 | 1.05E+08 |
| **jul** | -8.7540 | 5.37E+07 | -1.63E-07 | 1.000 | -1.05E+08 | 1.05E+08 |
| **jun** | -9.3002 | 5.38E+07 | -1.73E-07 | 1.000 | -1.05E+08 | 1.05E+08 |
| **mar** | -9.4847 | 5.46E+07 | -1.74E-07 | 1.000 | -1.07E+08 | 1.07E+08 |
| **nov** | -10.1436 | 5.43E+07 | -1.87E-07 | 1.000 | -1.06E+08 | 1.06E+08 |
| **dec** | -7.6052 | 5.44E+07 | -1.4E-07 | 1.000 | -1.07E+08 | 1.07E+08 |
| **apr** | -10.0632 | 5.44E+07 | -1.85E-07 | 1.000 | -1.07E+08 | 1.07E+08 |
| **aug** | -10.0374 | 5.45E+07 | -1.84E-07 | 1.000 | -1.07E+08 | 1.07E+08 |
| **may** | -9.4253 | 5.45E+07 | -1.73E-07 | 1.000 | -1.07E+08 | 1.07E+08 |
| **sep** | -9.2454 | 5.47E+07 | -1.69E-07 | 1.000 | -1.07E+08 | 1.07E+08 |
| **mon** | -18.6211 | nan | nan | nan | nan | nan |
| **tue** | -18.7379 | nan | nan | nan | nan | nan |
| **fri** | -18.5651 | nan | nan | nan | nan | nan |
| **thu** | -18.5239 | nan | nan | nan | nan | nan |
| **wed** | -18.4458 | nan | nan | nan | nan | nan |
| **previous\_campaign\_failure** | -31.4200 | nan | nan | nan | nan | nan |
| **previous\_campaign\_nonexistent** | -30.9942 | nan | nan | nan | nan | nan |
| **previous\_campaign\_success** | -30.4603 | nan | nan | nan | nan | na |

