Kyle Calabria

04/12/2018

Term Deposit Marketing Subscription Prediction

**Executive Summary**

This project is based off a white paper titled: A data-driven approach to predict the success of bank tele-marketing. This paper shared a portion of the data from the study on UCI’s data repository for which I used to conduct my project. The goal of this project is to predict whether the Portuguese banking institution’s clients from the study would subscribe to long-term deposits via a telemarketing campaign. The data available was related to the customer demographic, previous telemarketing campaigns, other, and social and economic indicator data. The dataset presented imbalanced classes with 11.27% of the samples subscribing to the term deposit. Given that the Portuguese bank in the study was under government pressure to increase its capital requirements, the primary consideration of the bank was maximizing its sales of term deposits. It was less concerned with disturbing clients and the costs associated with running telemarketing campaigns. For this reason, I elected to assess the success of the model predictions using the area under the Receiver Operating Characteristic curve which best represents the mission of maximizing sales. Had the bank been concerned with not disturbing clientele or reducing marketing costs, I would have placed higher consideration on the precision recall curve. I tested five types of models: Logistic Regression, Random Forest Classifier, Multi-layer Perceptron Classifier, K-nearest Neighbors, and a Support Vector Classifier. I also tested three different datasets which included the raw data, rebalanced data using upsampling and a data frame with re-binned features. The best model was a Random Forest Classifier fit on the raw data. The optimal parameters were 360 estimators, a minimum sample leaf of 4, auto max features and a max depth of 15. This models area under the ROC curve and test accuracy were 0.7906 and 90.14% respectively . The non-upsampled dataset outperformed the upsampled dataset which can be explained by the fact that decision trees are robust against imbalanced datasets. Three of the five most important features of the dataset were all economic indicators which suggested that the current economic climate plays a more significant role in predicting telemarketing campaign success that customer-specific data. Aside from these, the most important customer-specific features were the number of days that passed since the client was last contacted from the previous campaign, age, and the number of contacts performed before this campaign.

**Problem**

The problem statement for this project was based on a white paper titled: A data-driven approach to predict the success of bank telemarketing. The goal of this project is to predict whether a Portuguese banking institution’s clients will subscribe to their term deposit product based on information obtained prior to the sales call. 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. The original experiment was conducted on data from a Portuguese banking institution between the times of May 2008 and June 2013 which includes the time period during which the financial crisis occurs thus captures some of its effects in the data as well as the recovery process. Unfortunately more data from prior to the crash was not available which would have allowed establishment of a baseline of a pre-crisis norm. Although the goal of the model is to predict whether or not a client will subscribe to a term deposit, more useful to bank managers would likely be the probability that a client will subscribe as oppose to the prediction itself. A model producing probabilities allows a bank manager to set a threshold over which telemarketing agents will reach out. This allows a bank manager to reduce the cost associated with a telemarketing campaign by having to pay fewer telemarketing agents to reach out to fewer clients i.e. only those above a certain threshold of probability to convert. This additionally adds value to the business by not bothering clients which could potentially reduce the customers lifetime value to the bank. Having said that, the Portuguese bank in the original study was under pressure from the government to maximize its capital reserves post-financial crisis and thus was more concerned with maximizing sales as oppose to avoiding disturbing clients. Given this incentive, the bank favors more sensitive approach and the the area under the curve (AUC) of a Receiver Operating Characteristic (ROC) curve is a superior method for measuring model effectiveness to a Precision Recall (PR) curve. This is because the bank is more concerned with maximizing sales than with bother clients and thus precision is less of an important factor. We also care about the models generalizability to other populations which is better represented by an ROC curve. A second important benefit of the ROC curve is that the bank manager can optimize the true positive rate and false positive rate to match the business or branches needs. In the scenario that the client did care more about bothering customers the precision of predictions would be more significant and a PR curve would be more appropriate. For this hypothetical, I will include some PR curve and analysis in addition the ROC curve. In addition to producing the probability of conversion for each client with a model that maximizes the ROC curve, an equally important goal of the project is to determine which features are most important in determining conversion. Conversion features include features that are controllable by the bank i.e. frequency of phone call to customer and features that are uncontrollable by the bank i.e. Euriobor 3 month rate. For controllable features, the bank can attempt to optimize these for maximizing conversions. For Un-controllable features, the bank can coordinates its telemarketing campaign timing/investments with these to either mitigate or optimize their impact.

**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 target feature is imbalanced with 11.27% of samples being positive. This imbalance will produce challenges in the model development phase and will be address in the model building section. 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.

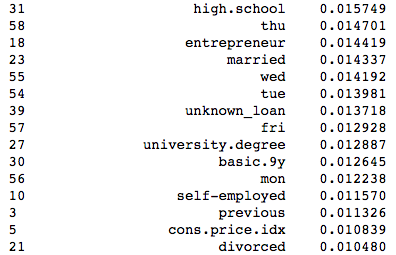
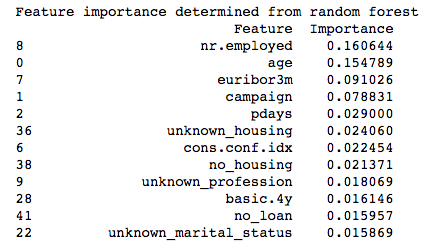
All of the numeric data is in a good format in raw form for training models on. I hot-one encoded the following categorical features: job category, marital status, education level, loan default status, housing loan status, generic loan status, phone type, day of the week of last contact, outcome of previous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

**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**

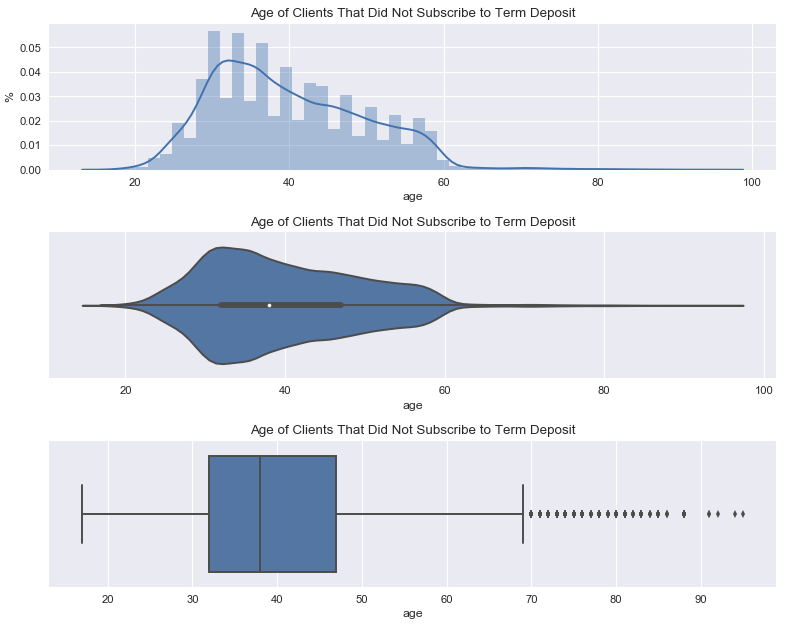
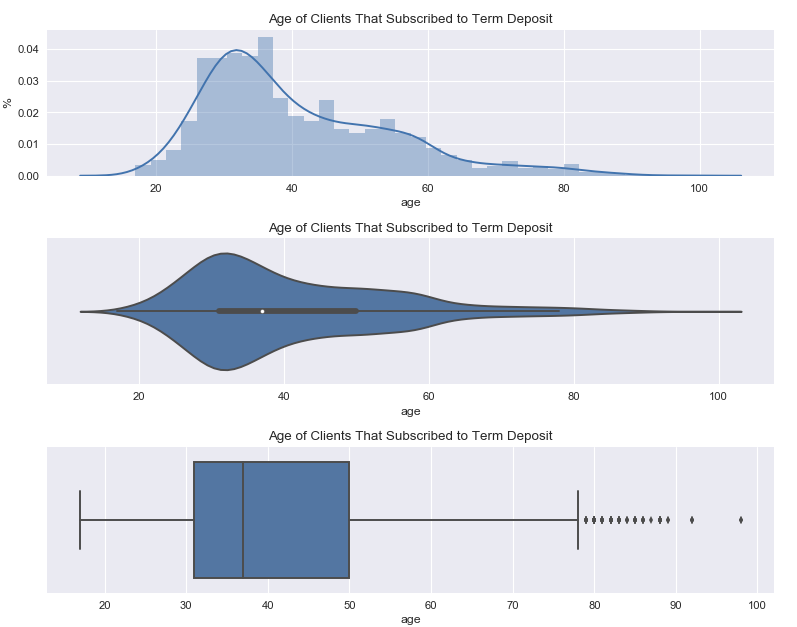
To orient myself of some of the important features to look at before beginning visual analysis of each individual feature, I ran a random forest model and created a feature importance data frame. Shown below are all of the features with a importance coefficient above .01. Social and economic context 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.

****

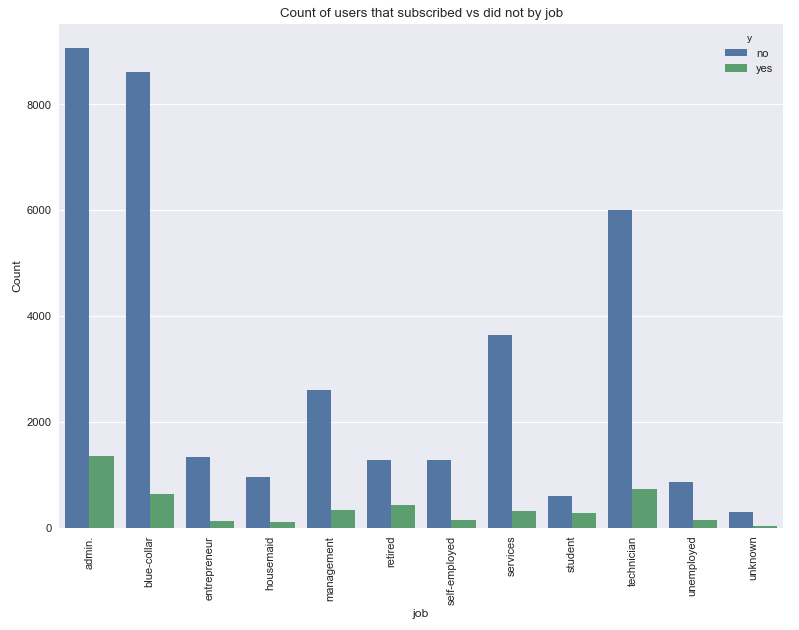
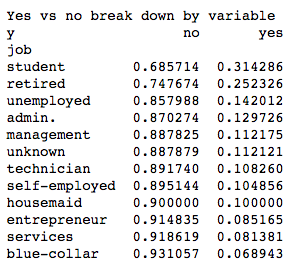
Age is the second most important feature which and the highest of all customer demographic information. This might be due to the fact that it is the only numeric feature of all customer demographic information thus all of the feature’s informational value is contained in a single column, whereas a feature 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.

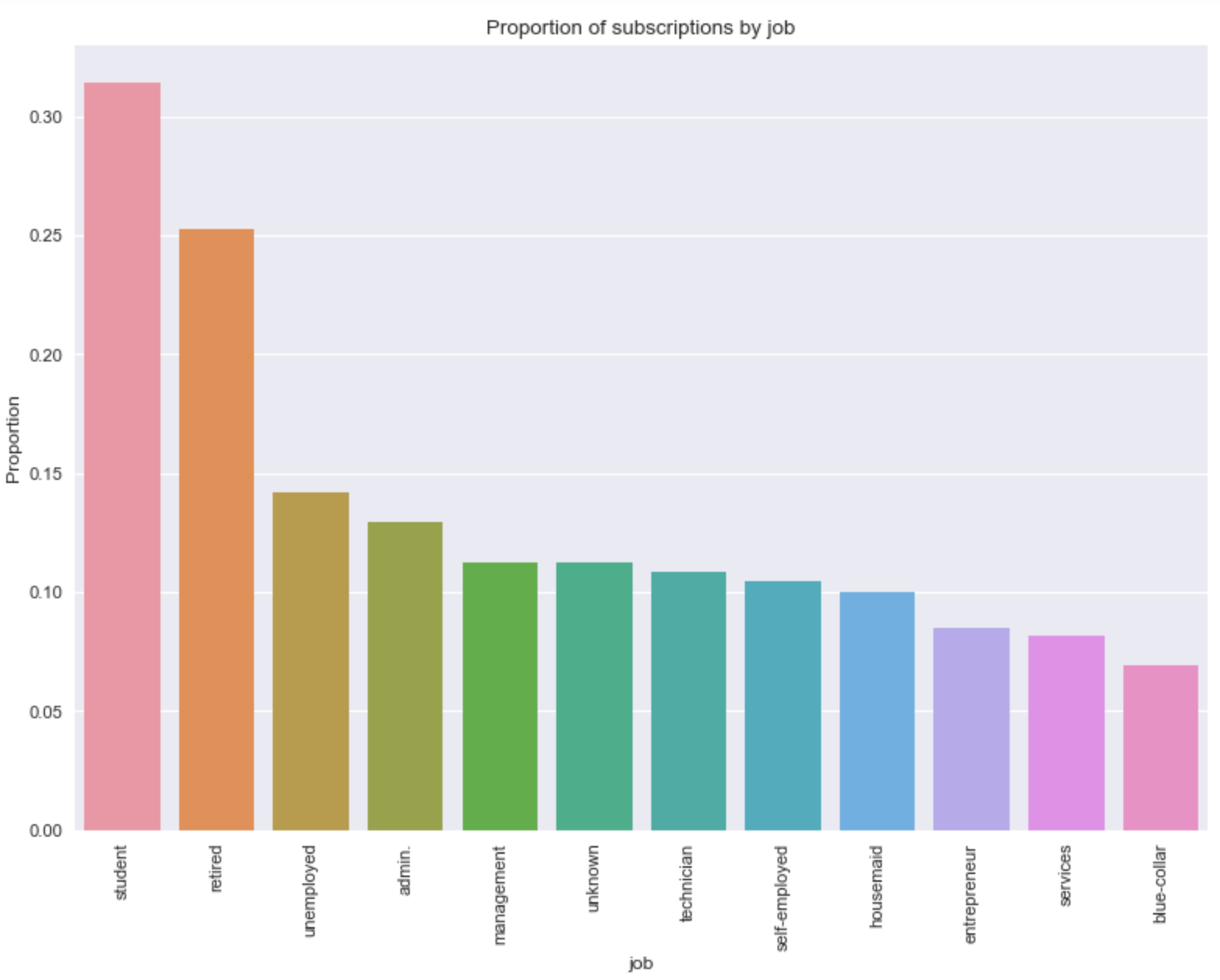
Of the campaign features, the number of contacts during the campaign was the most important followed directly by the number of previous days since the client was contacted in a previous campaign. I assume there is an optimal range of number of contacts that it would take to persuade a client to the term 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.

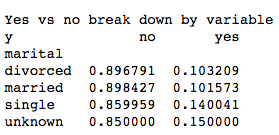
The above analysis and results are limited by the fact that they are specifically relevant to the model and don’t necessarily apply to other models or at an absolute level. Having said that, the feature importance is still likely to give me a rough idea of key features to explore further.



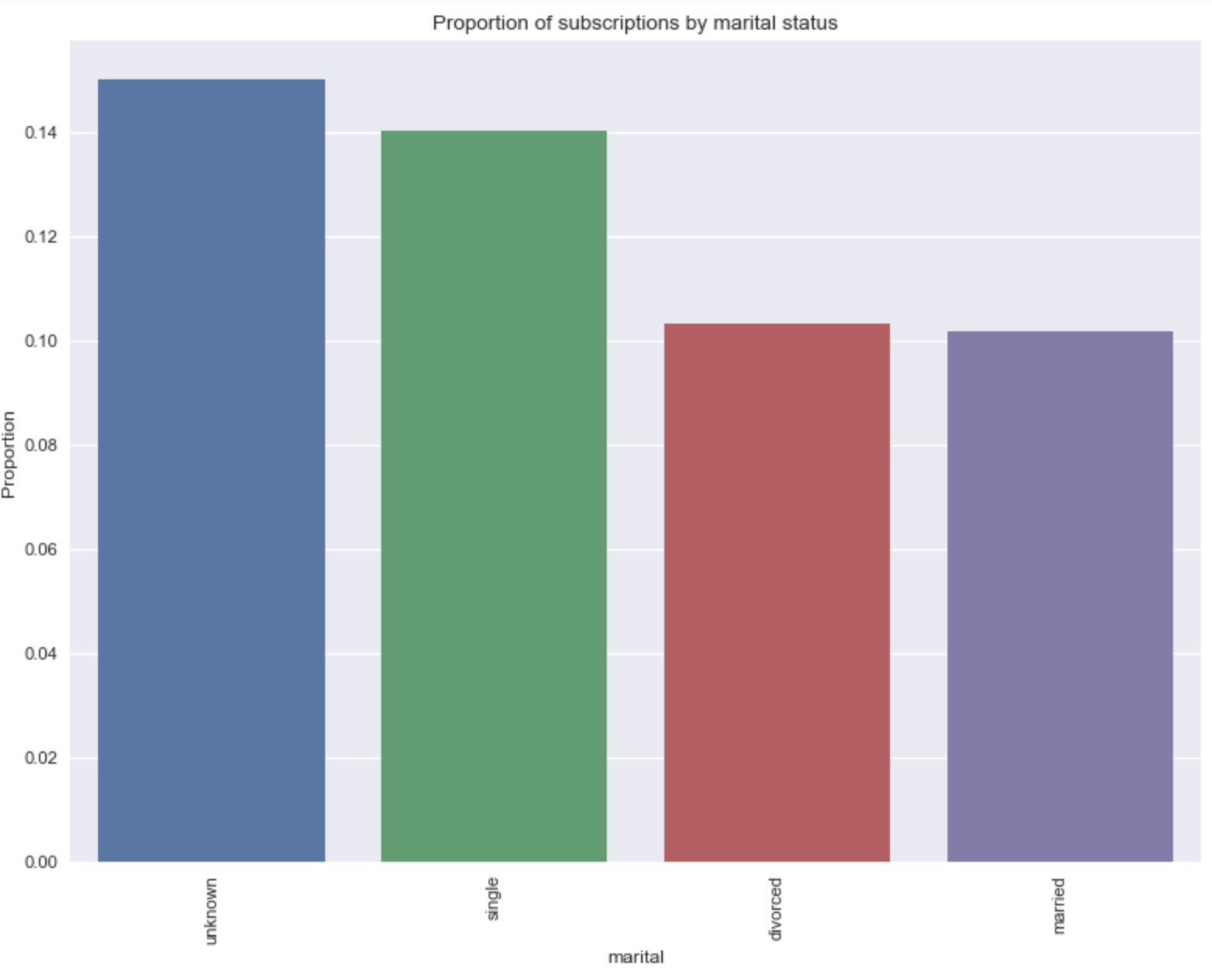
The counts by age for clients that subscribed vs those that did not follow different distributions. Clients that did subscribe follow a smoother distribution whereas clients that did not has a much more jagged distribution. The count of those that did not subscribe drops off sharply at 60 years of age 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.



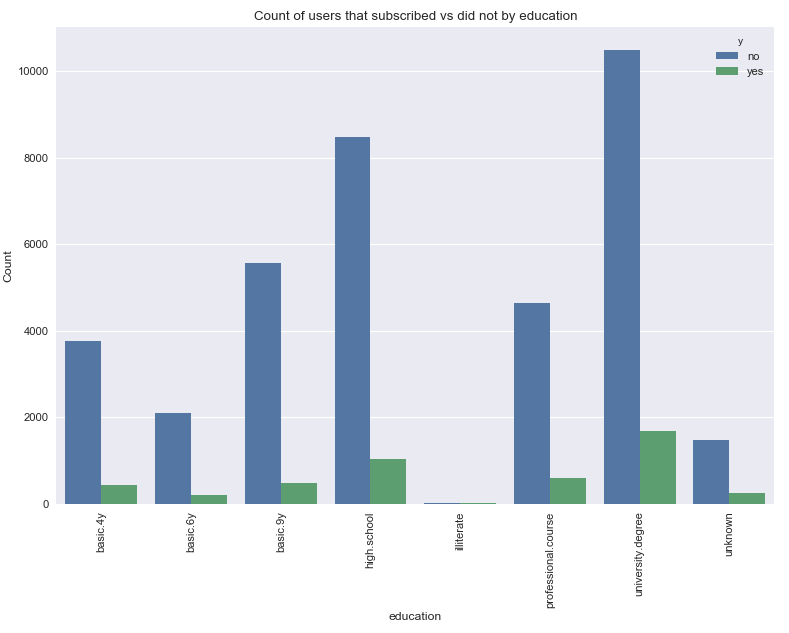
Students, retirees and unemployed have the highest proportion of subscriptions amongst professions, but only makeup a small percentage of the population. The fourth highest proportion of subscriptions amongst professions is admin., but this profession contains the largest percentage of the population.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

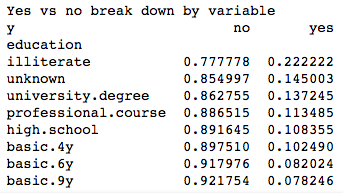
**Marital Status**

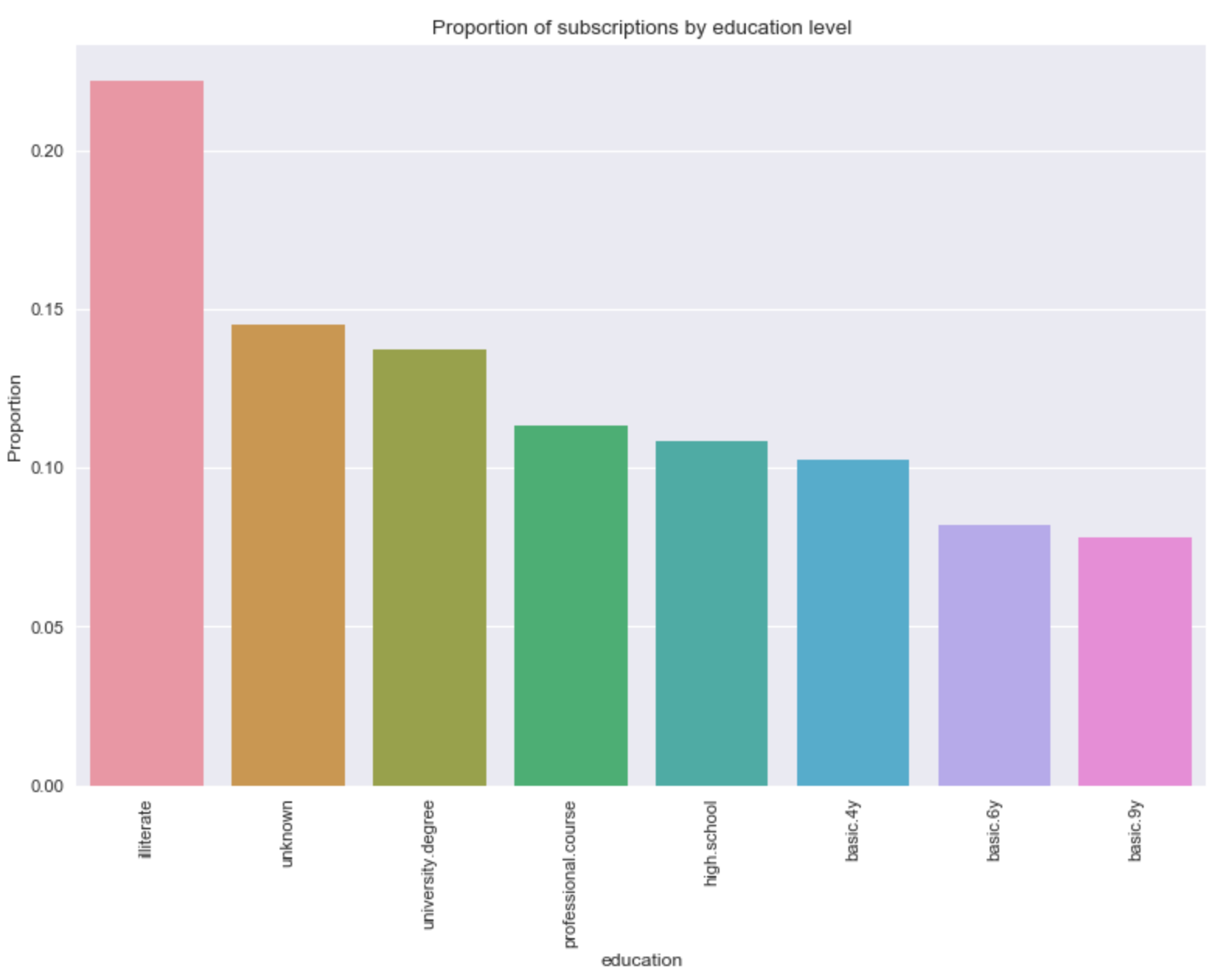
For material status there appears to be two groupings by proportion of clients to subscribe to the term deposit: those that have been married i.e. married and divorced and those that are either single or unknown. This pattern might indicate higher costs associated with raising a family resulting in the 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.

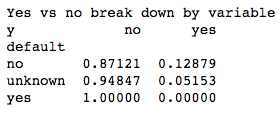
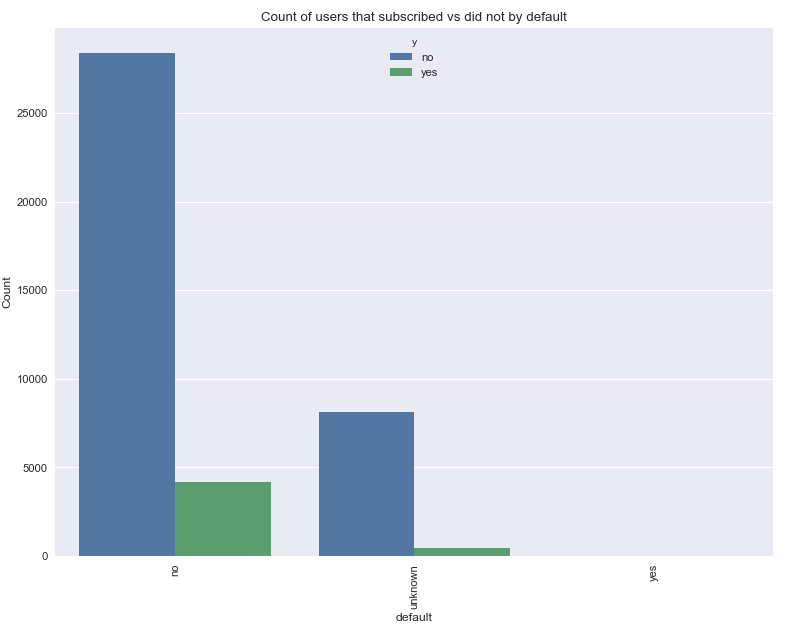


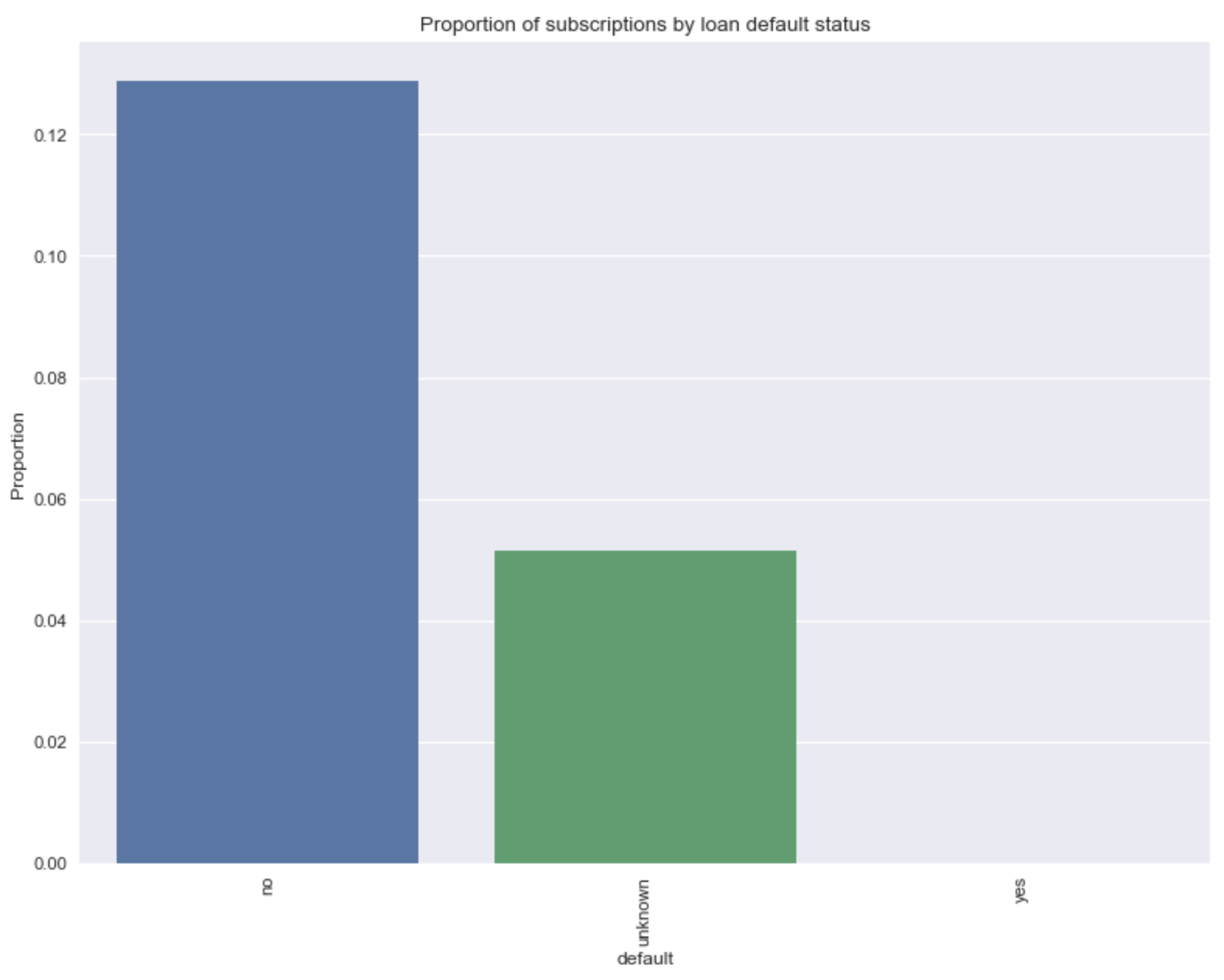
**Education**

****

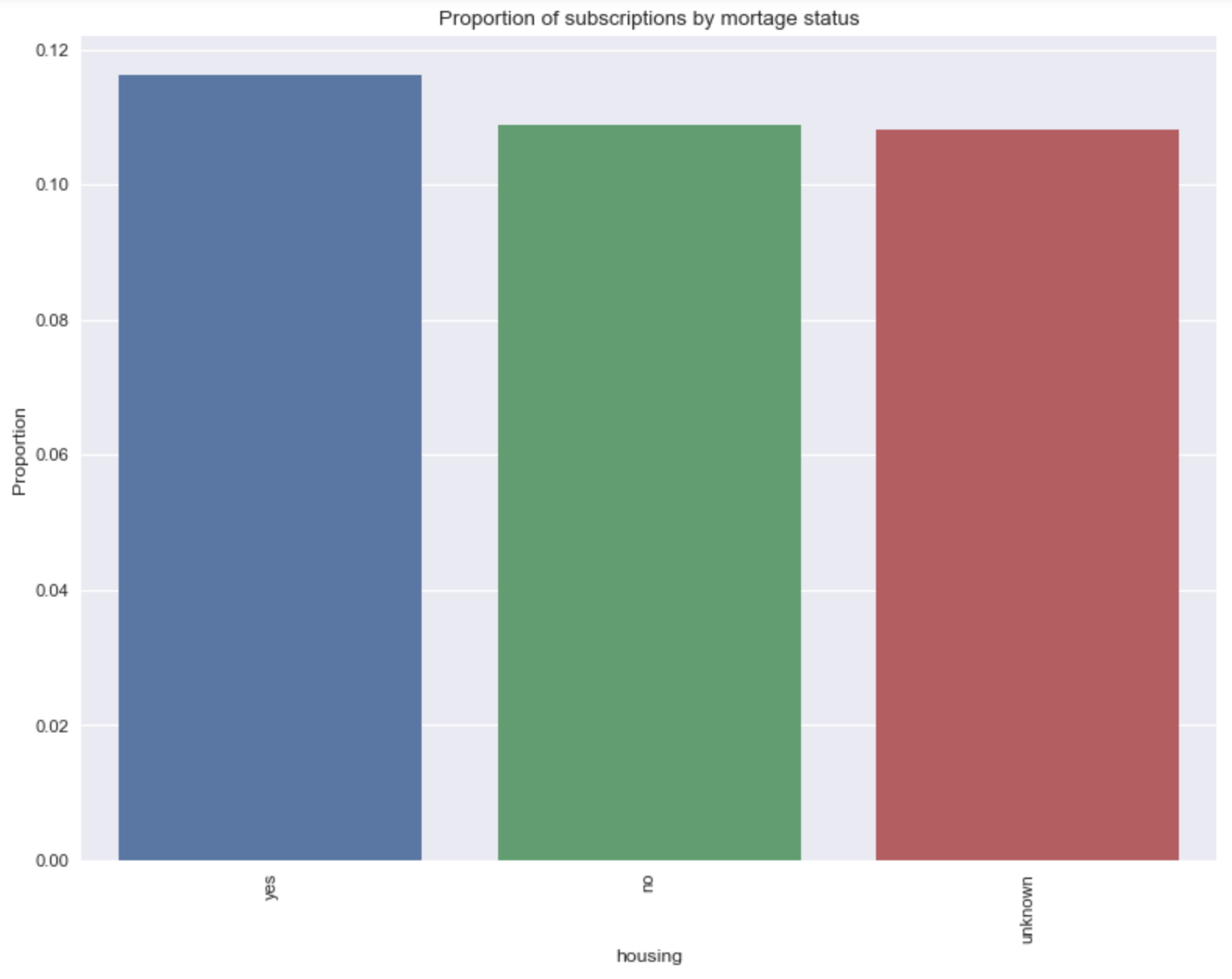
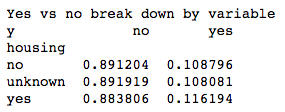
****

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**

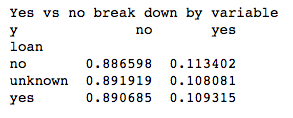
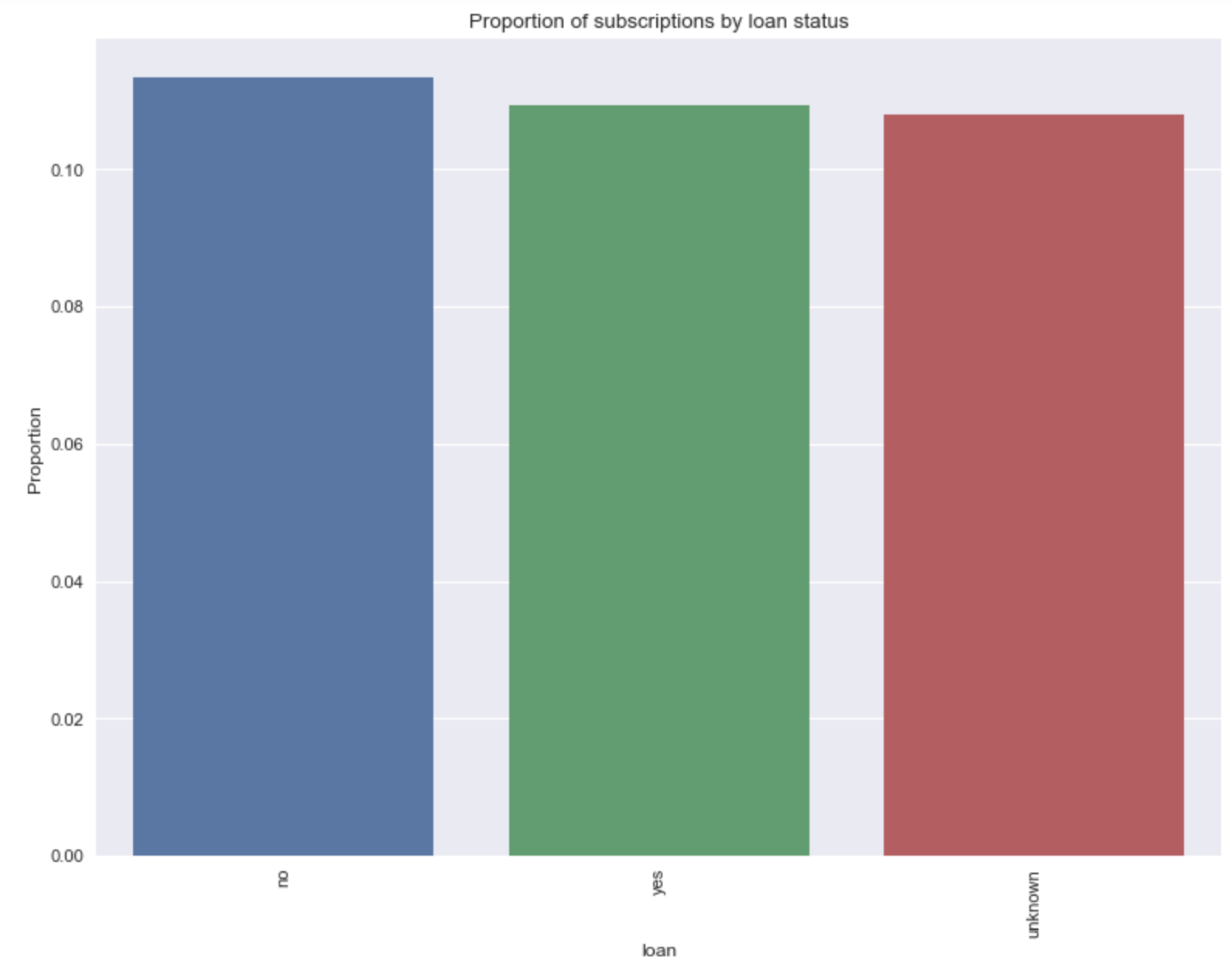
Zero clients that are in default were investing in the term deposit which makes it a highly predictive variable; however, clients in default only makes up a negligible percentage of the population. There is also a significant margin between the proportions of clients known to not be in default vs clients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

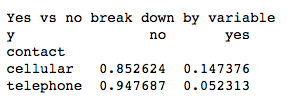
**Mortgage Status**

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.

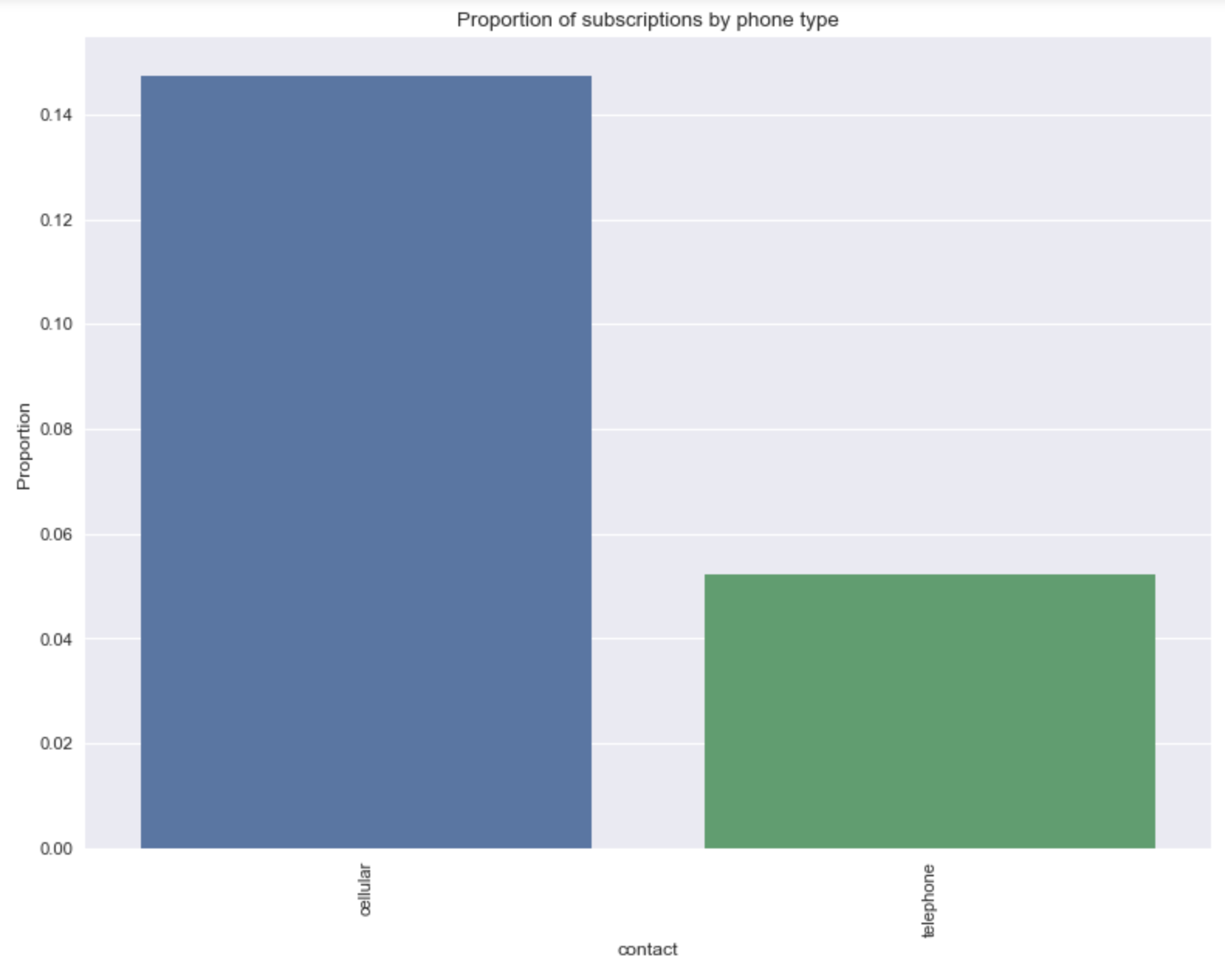
**Loan Status**

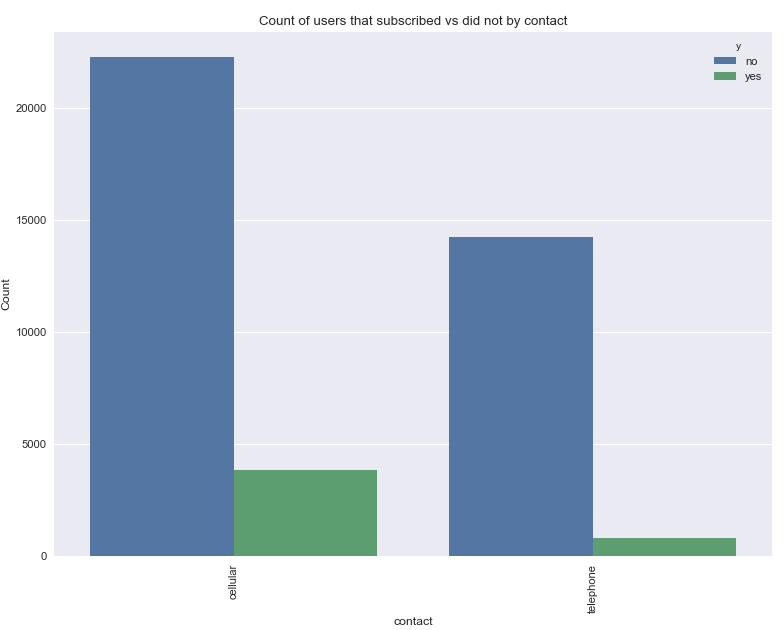
Loan status also had little correlation on the proportion of clients within the category to subscribe to the term deposit.

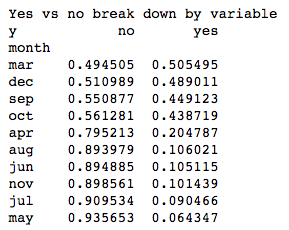
****



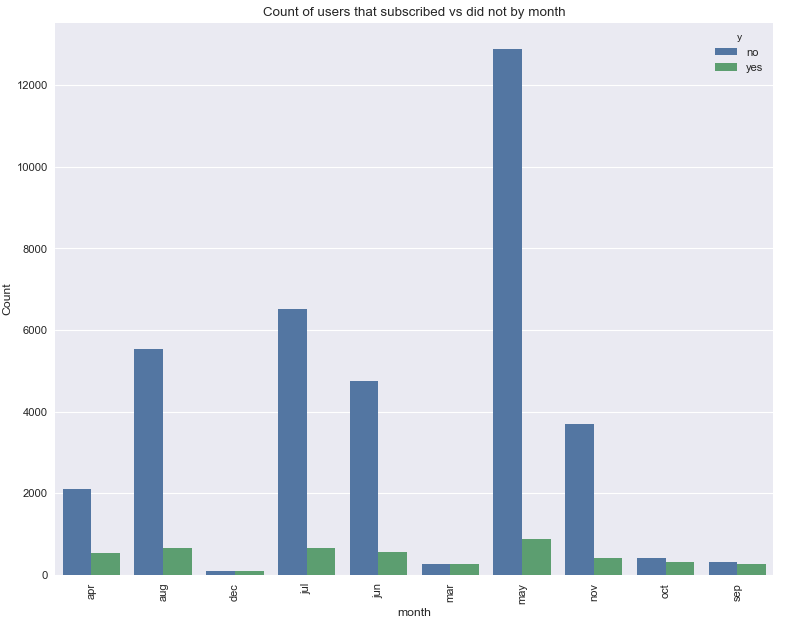
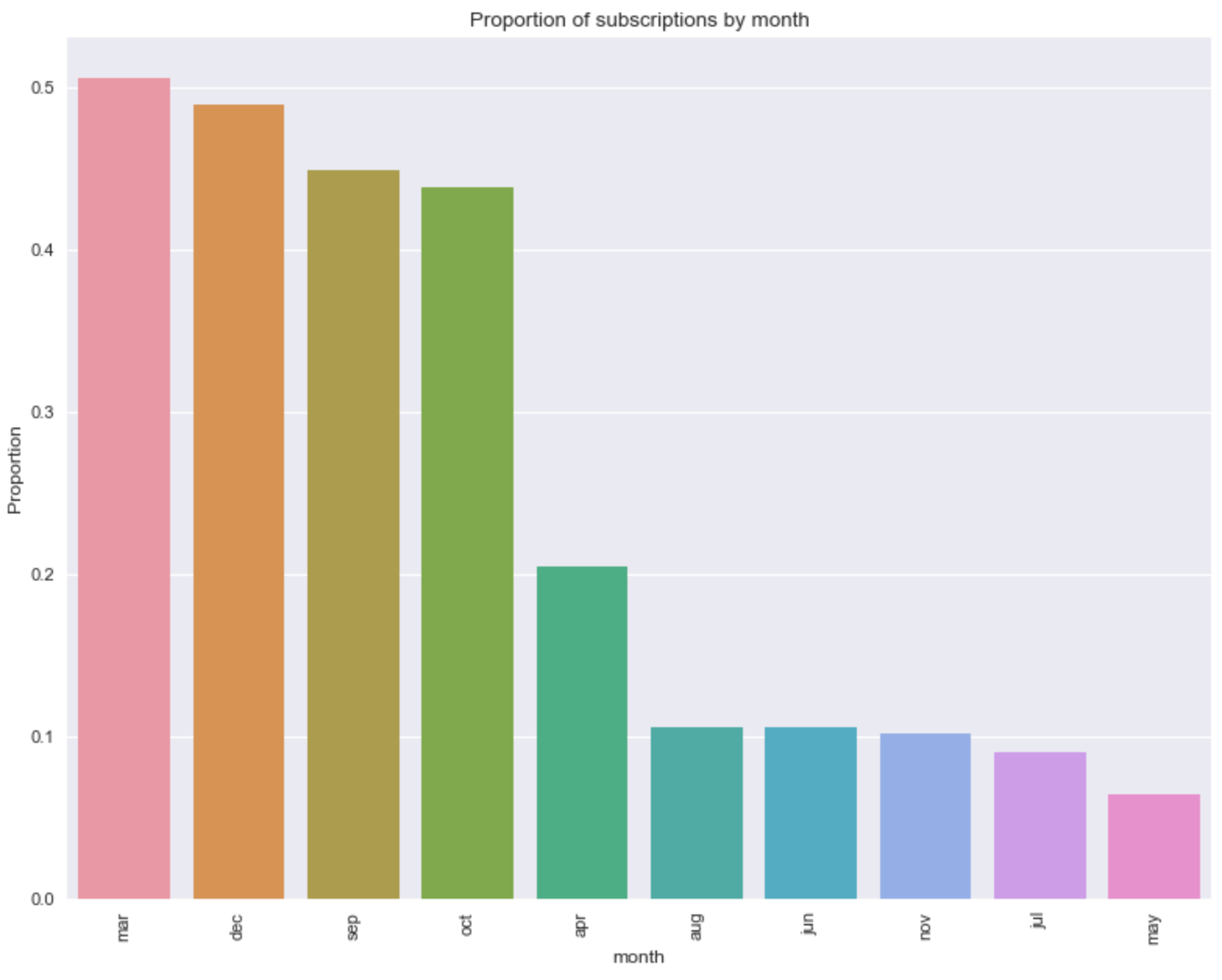
**Phone Type**

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.

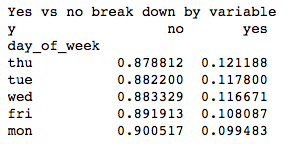
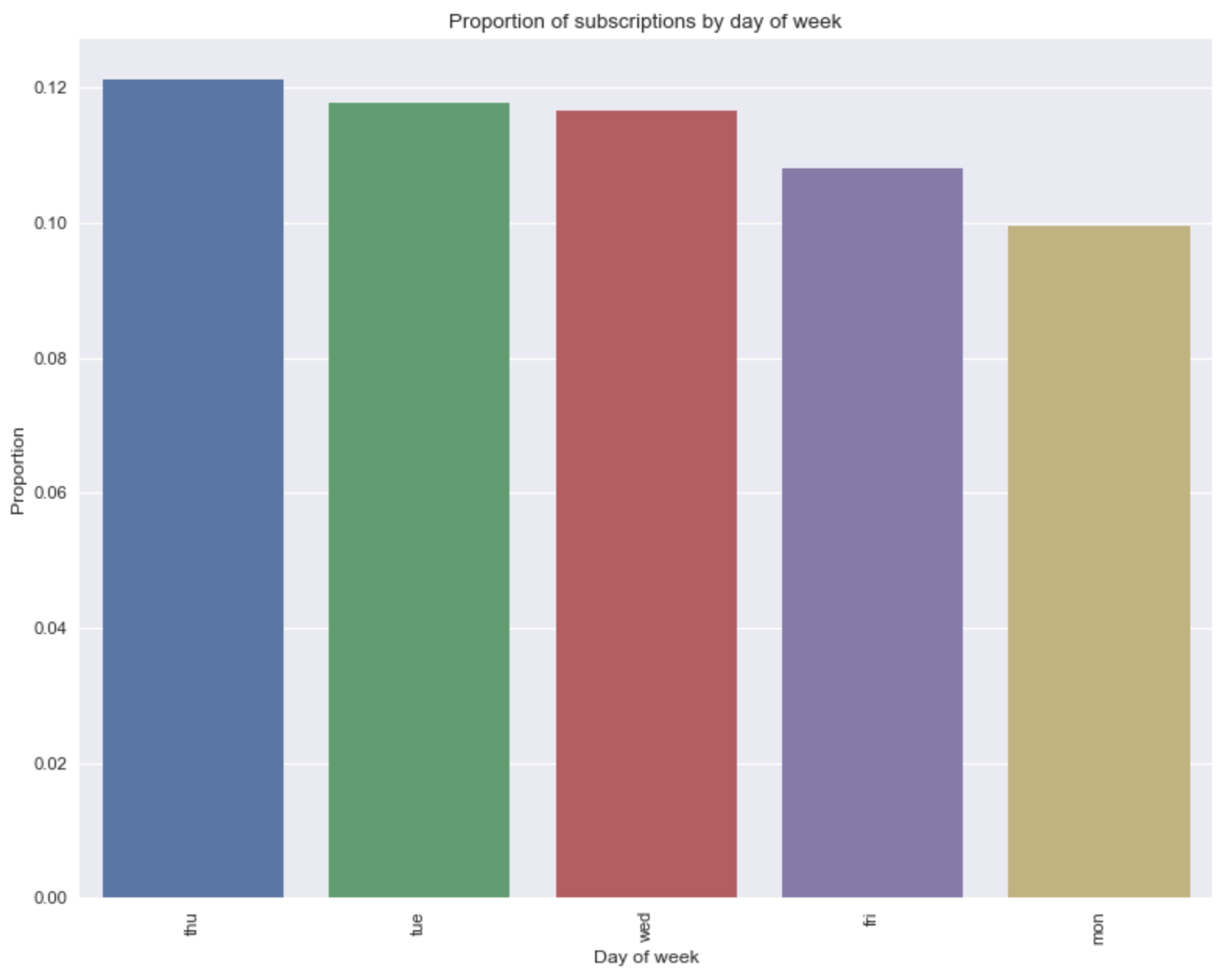


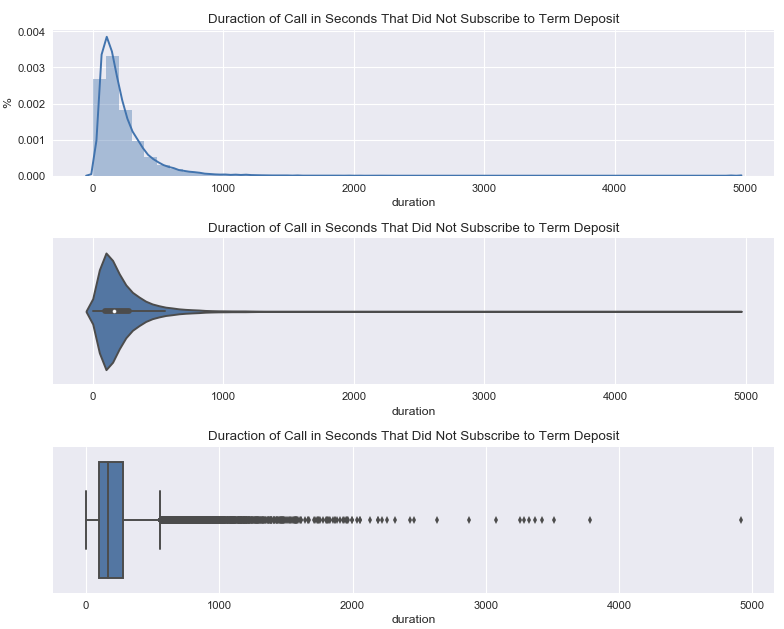
**Month**

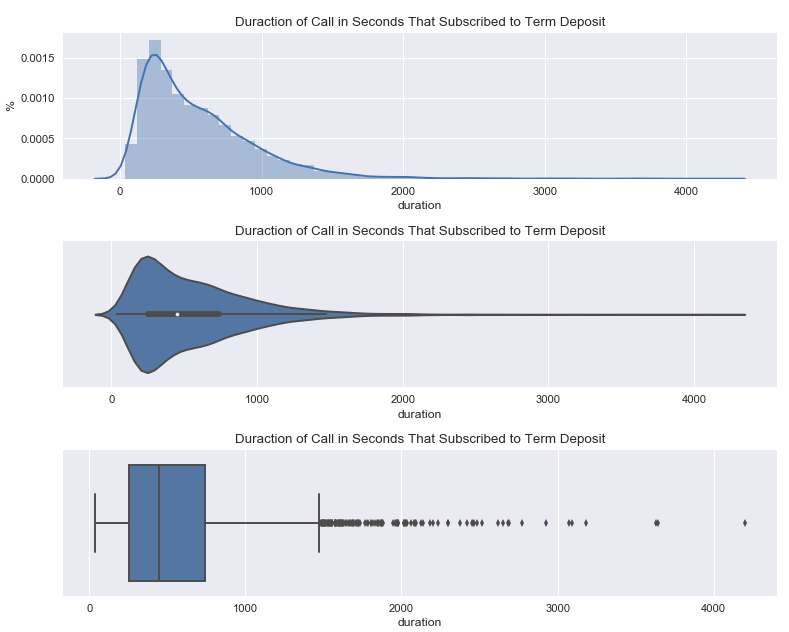
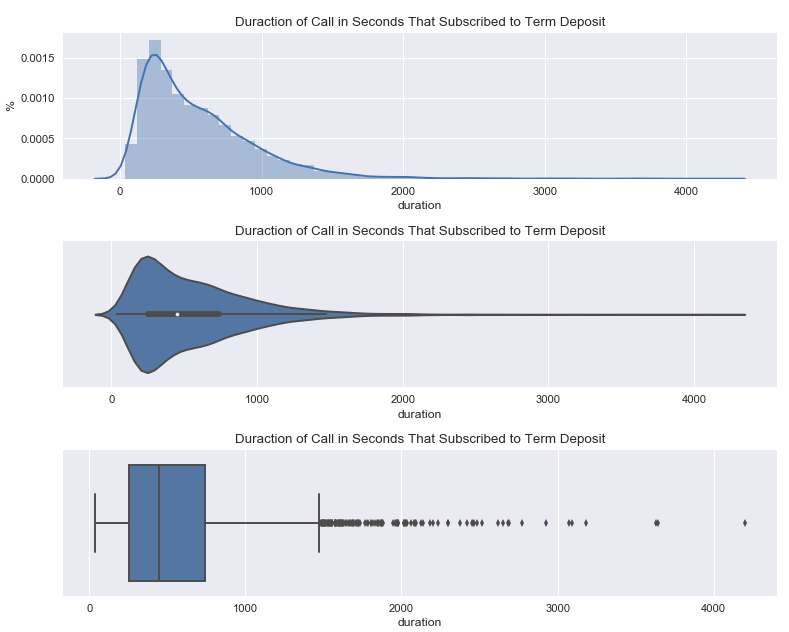
The month during the call was highly correlated with whether the client subscribed or not. March, December, September and October all had proportions of subscription above .4 whereas all other months were under .21. Given these proportions alone, someone might assume that the proportion of subscription 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.



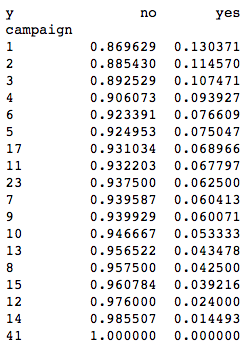
**Day of Week**

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.

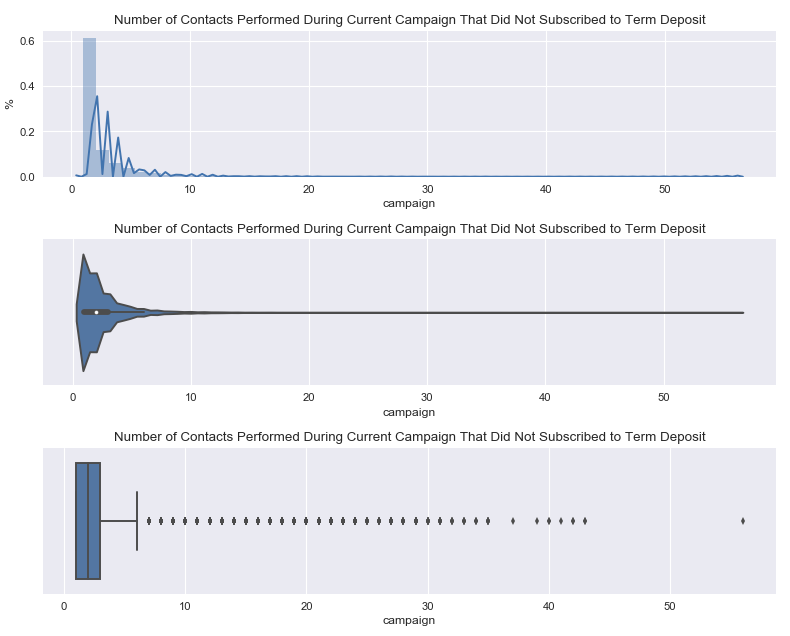
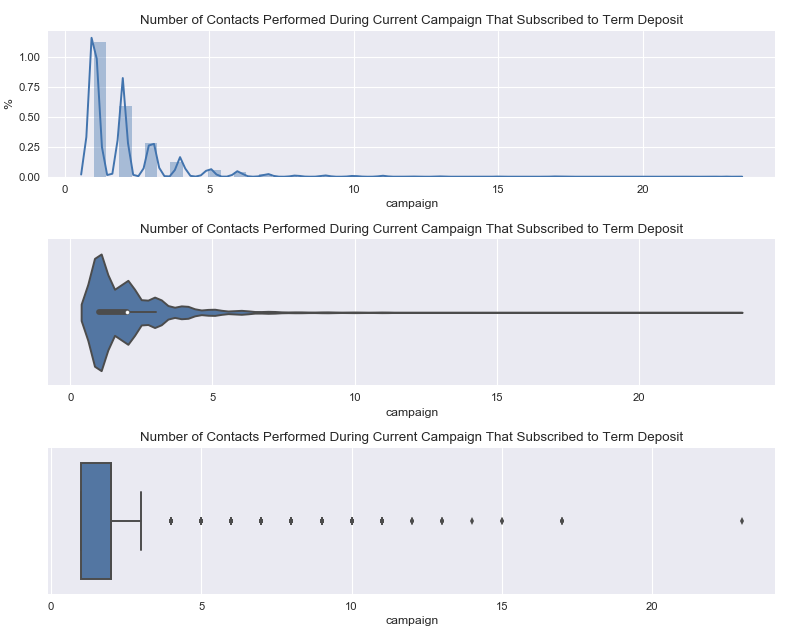
**Duration of Final Phone Call**

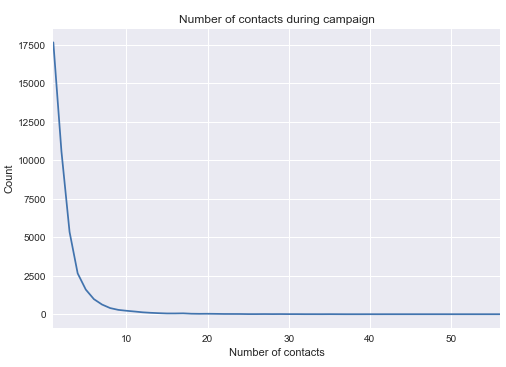
****

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.

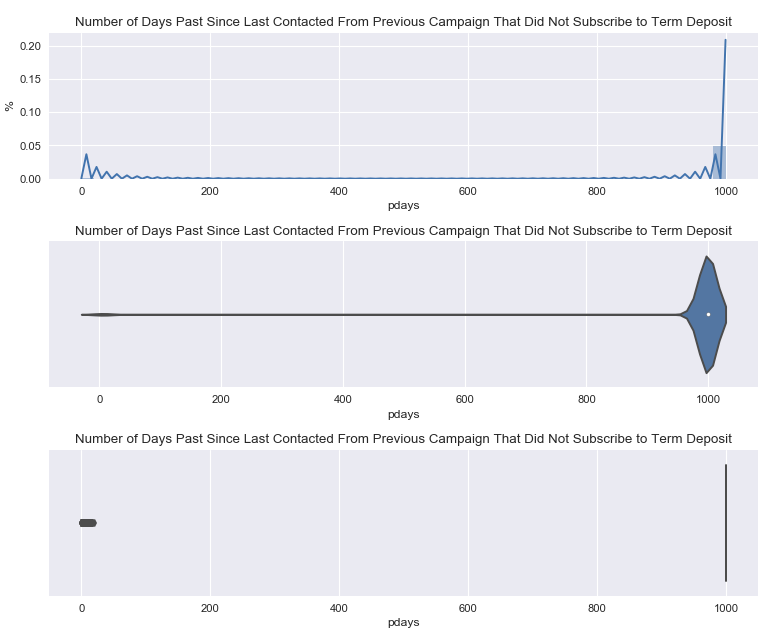
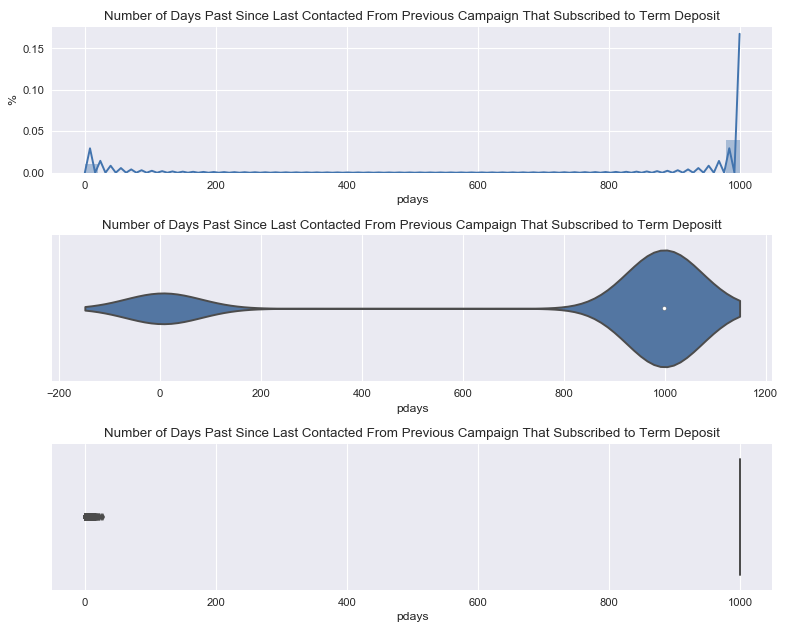
**Number of Contacts During Campaign**

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%.

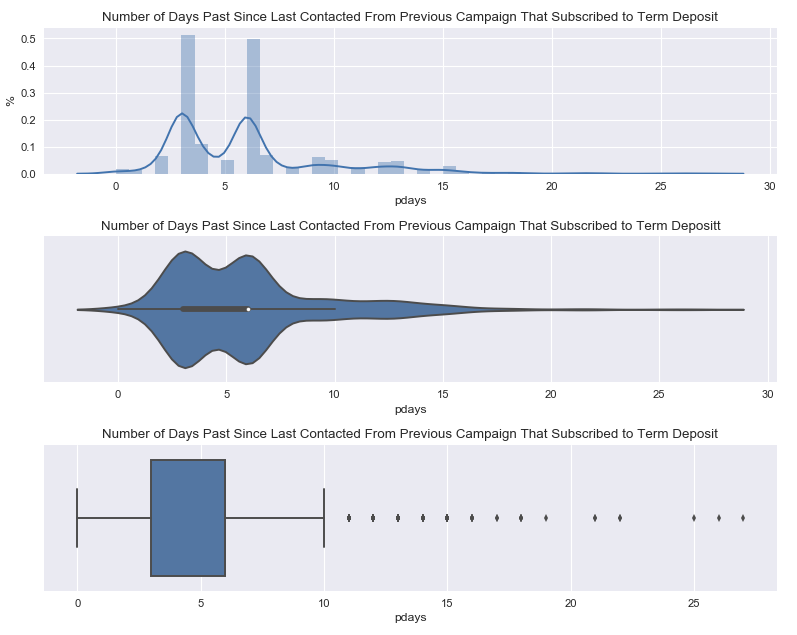
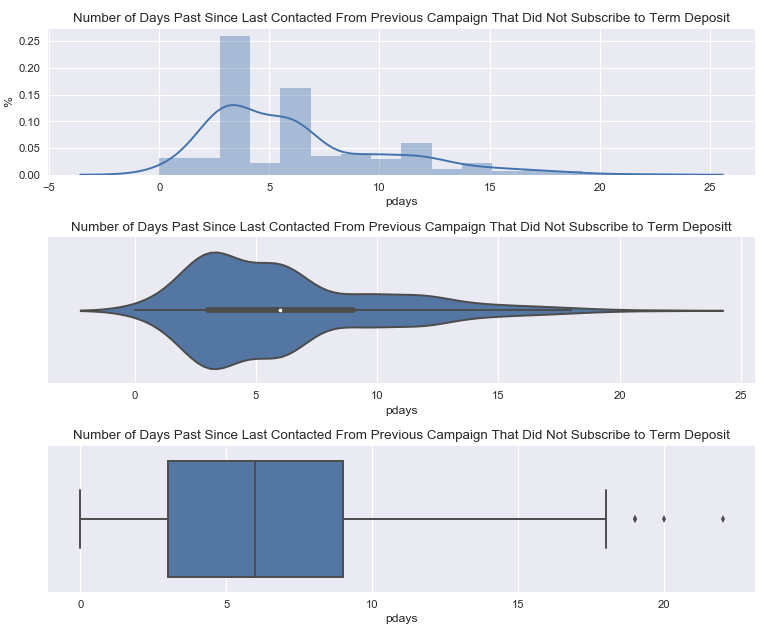


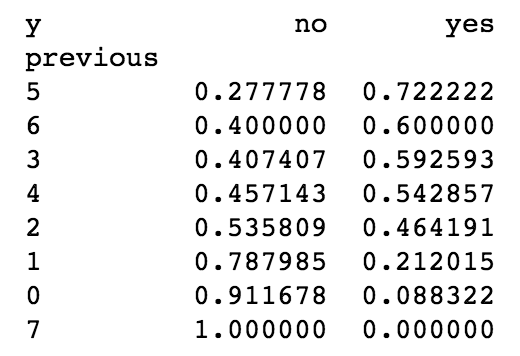


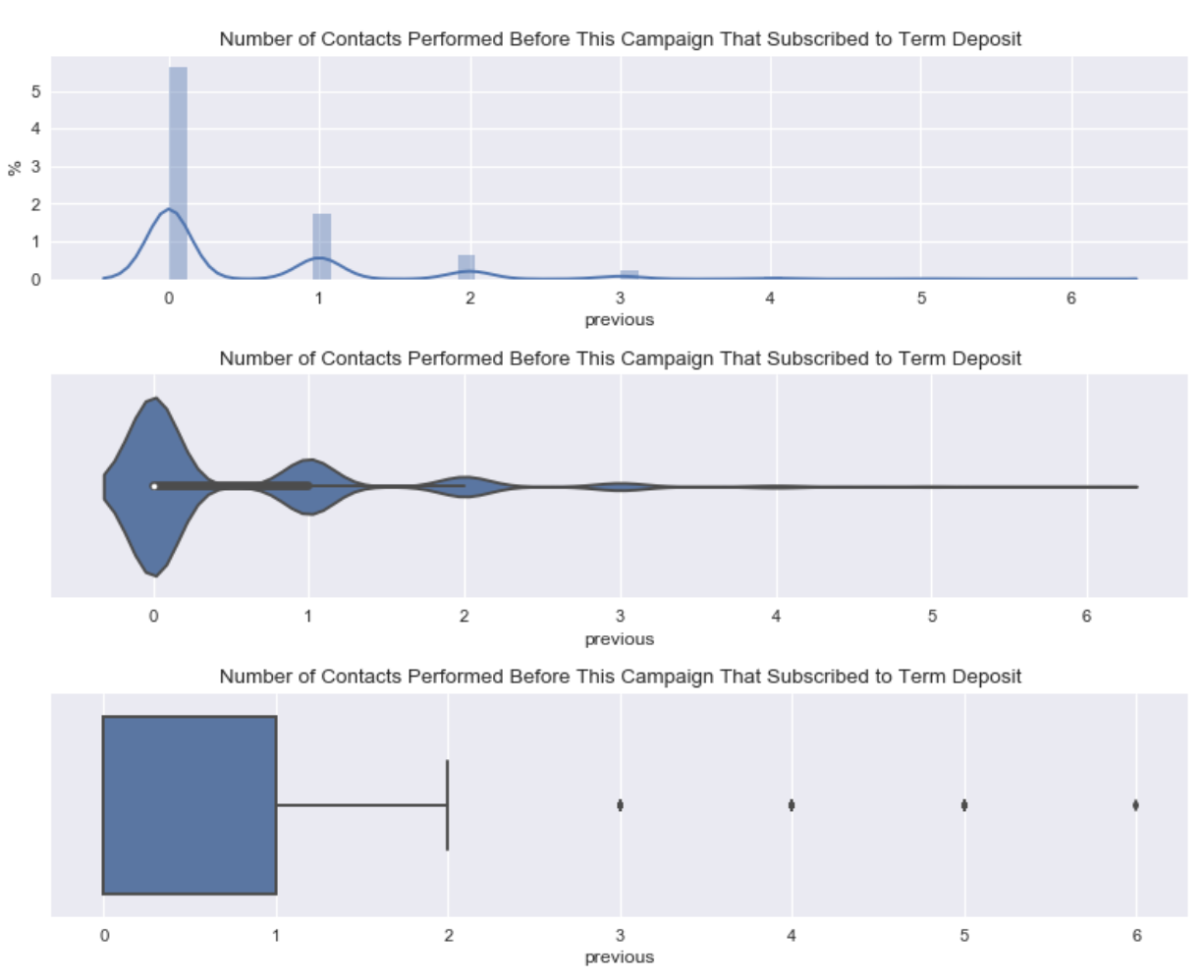
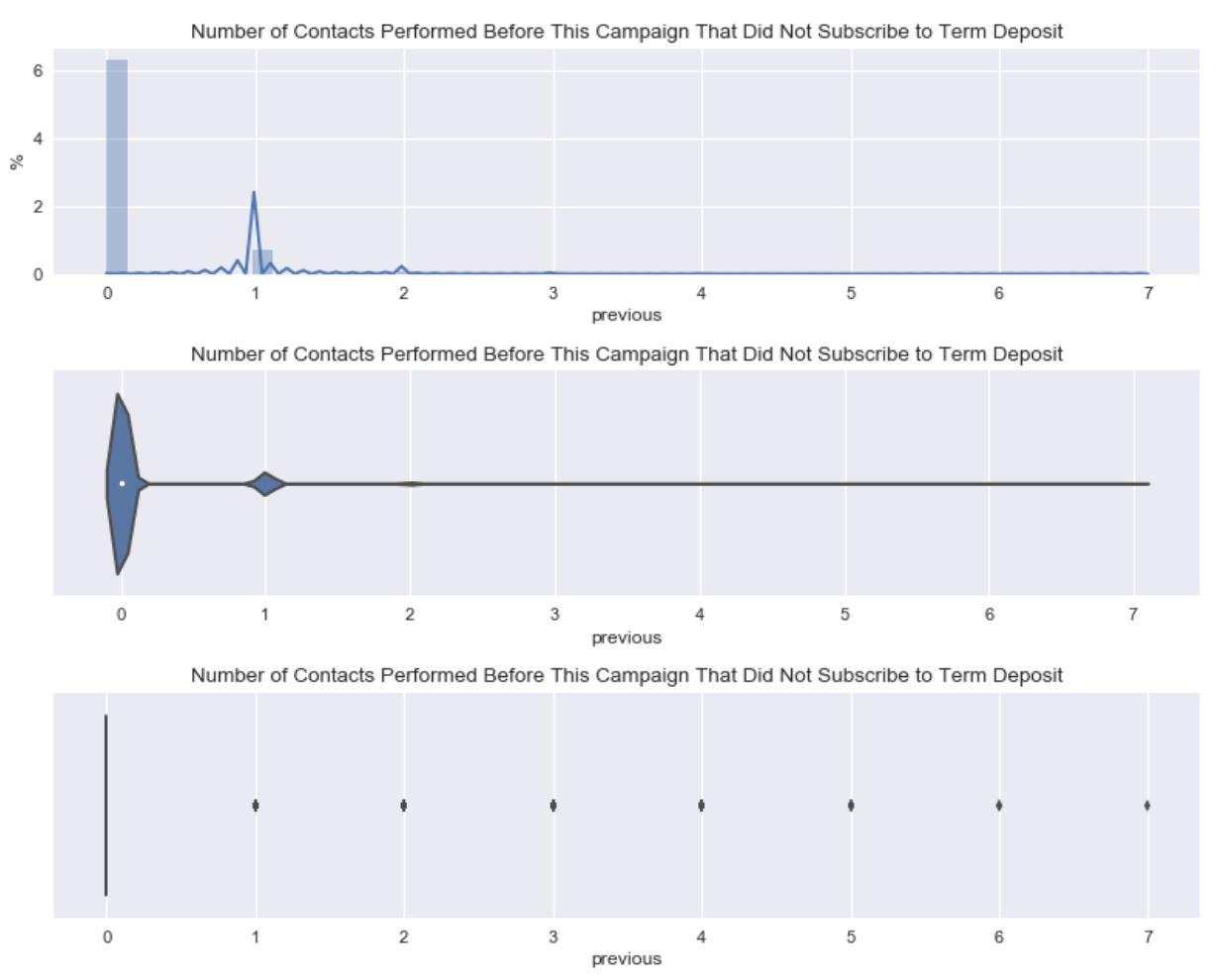
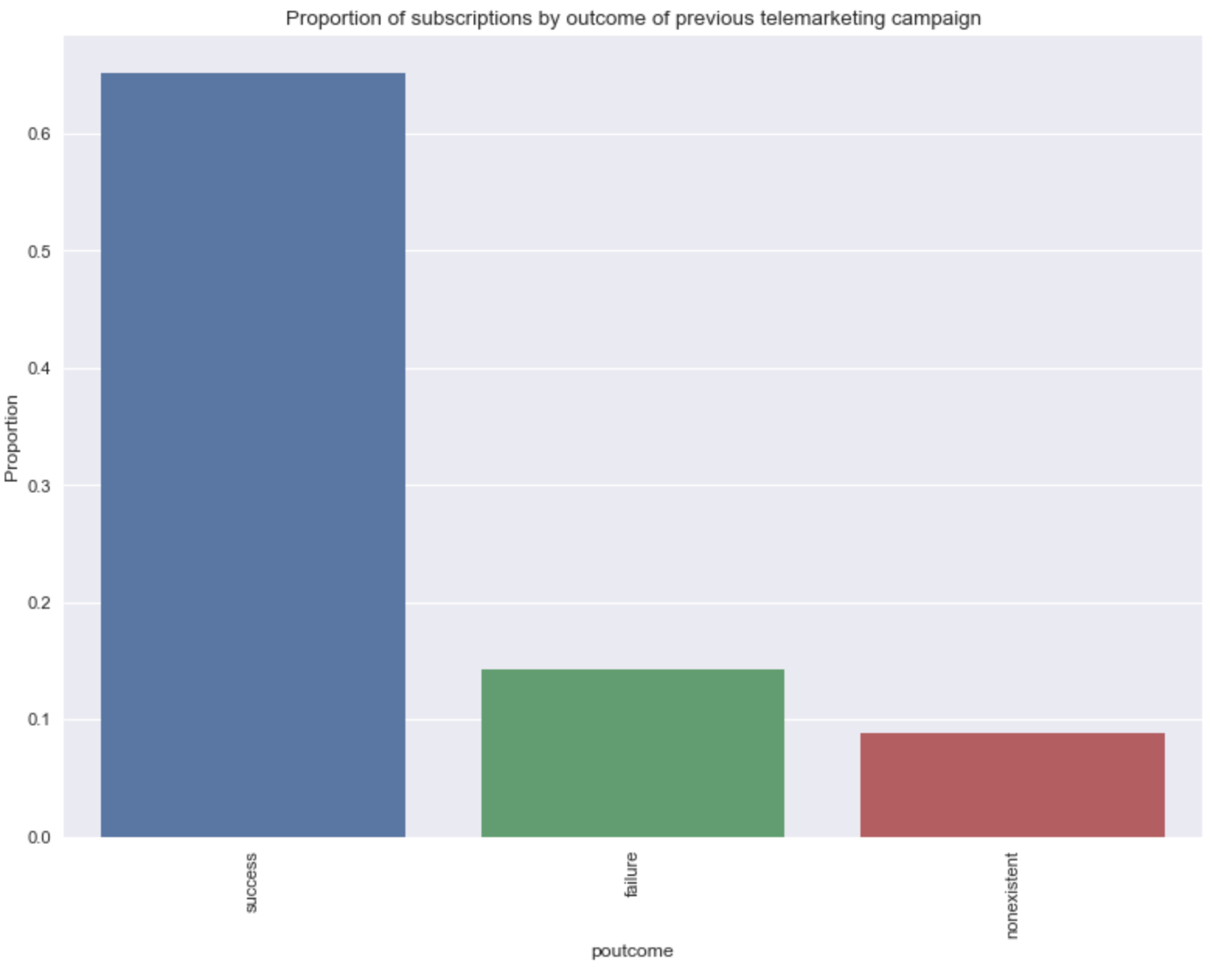
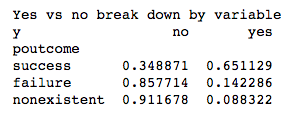
**Number of days since client was last contact from previous campaign**

****

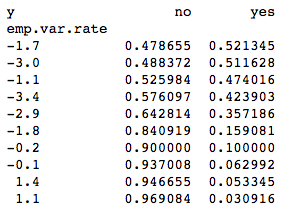
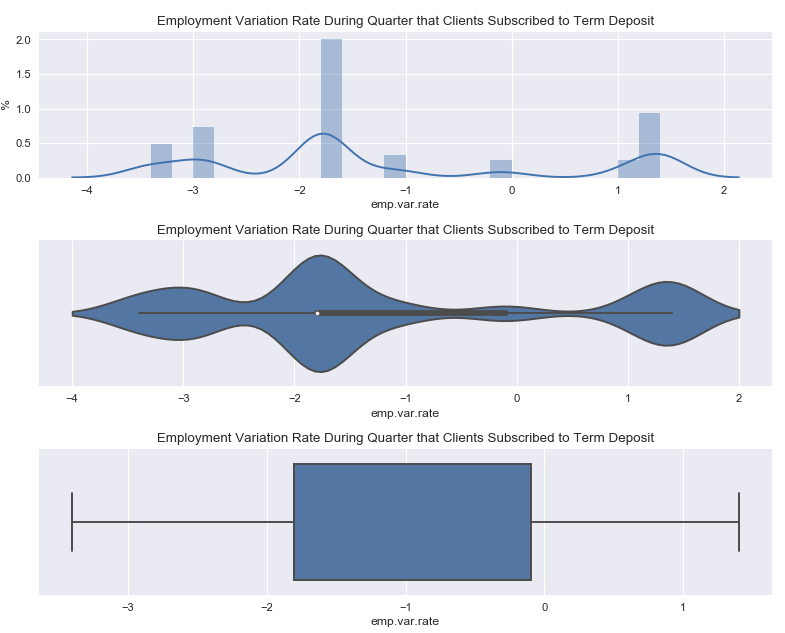
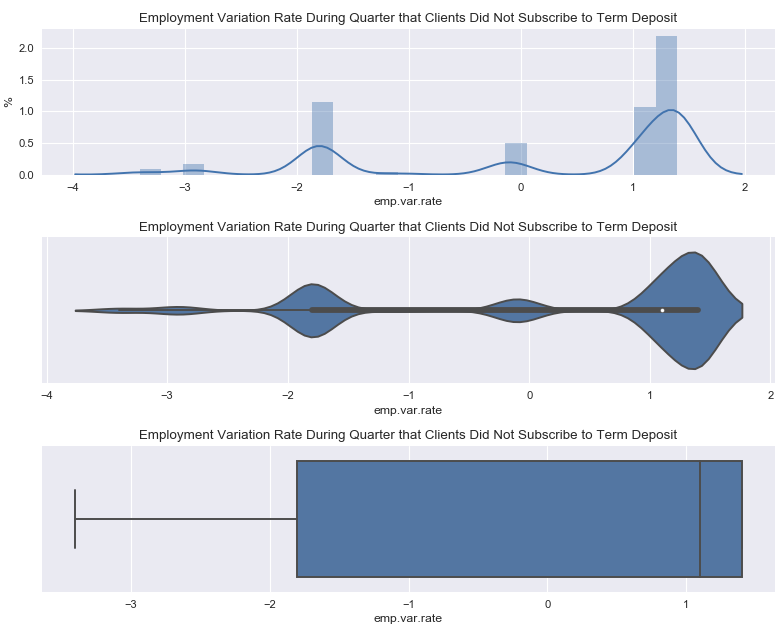
From the above plots it is evident that the distribution is lopsided due to all of the clients that have not previously been contacted which is encoded as 999. Below are the same plots with the clients that have not been contacted removed. Due to the fact that 96% of the data contained values of 999 representing no previous contact, the visualizations directly above provide little insight for the entire dataset.



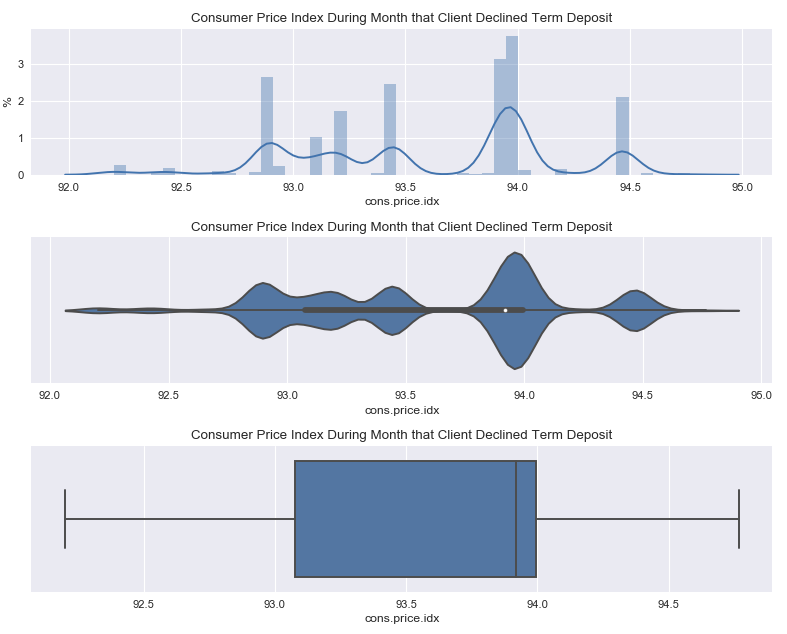
**Number of Contacts Performed Before this Campaign**

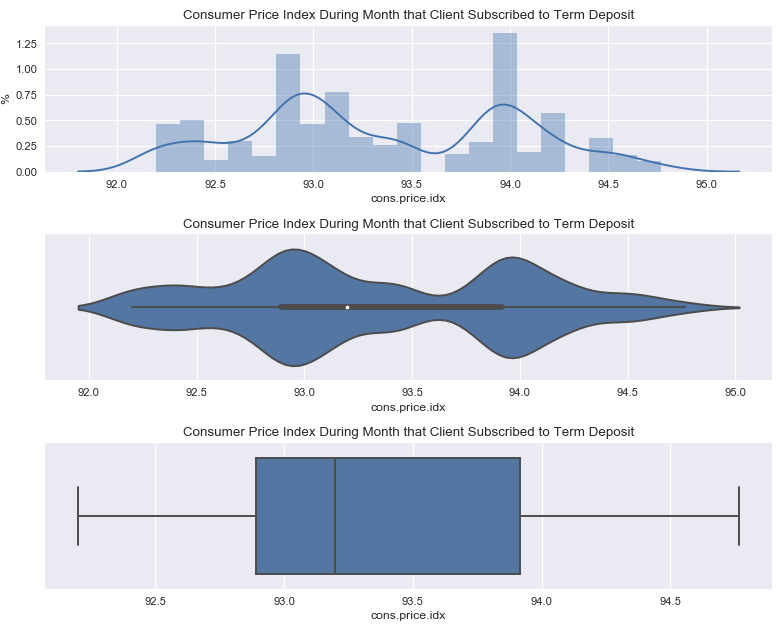
**Outcome of Previous Tele-marketing Campaign**

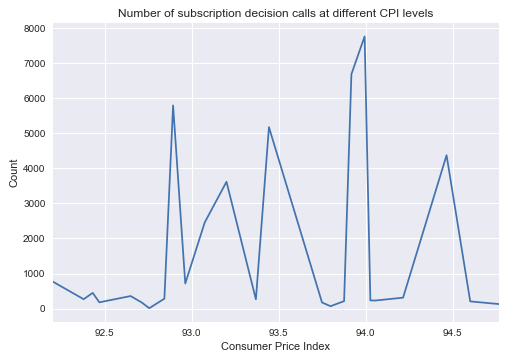
The proportion of term deposit subscriptions for clients having subscribed to a term deposit in a previous marketing campaign is four times higher than those that did not sign up in a previous campaign and over six times higher than had the client not been marketed a term deposit in the past at all.

**Employee Variation Rate**

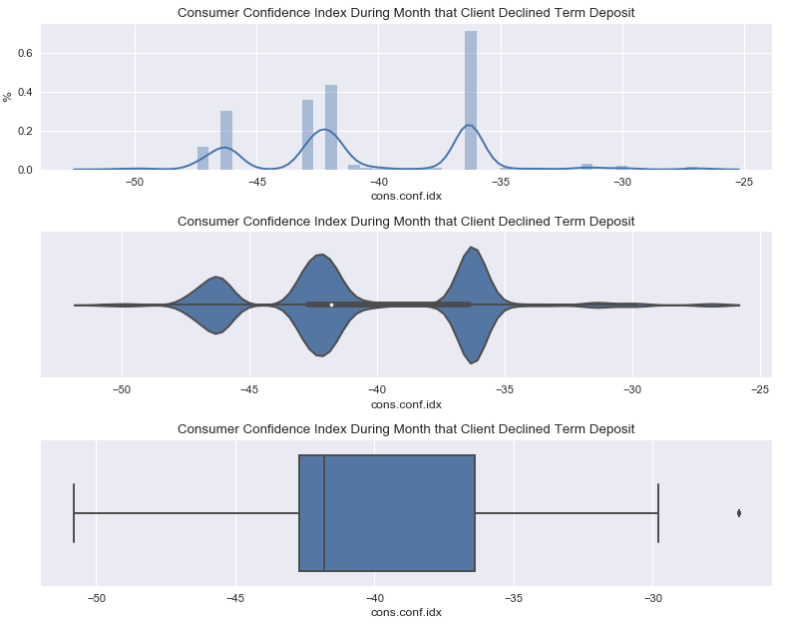
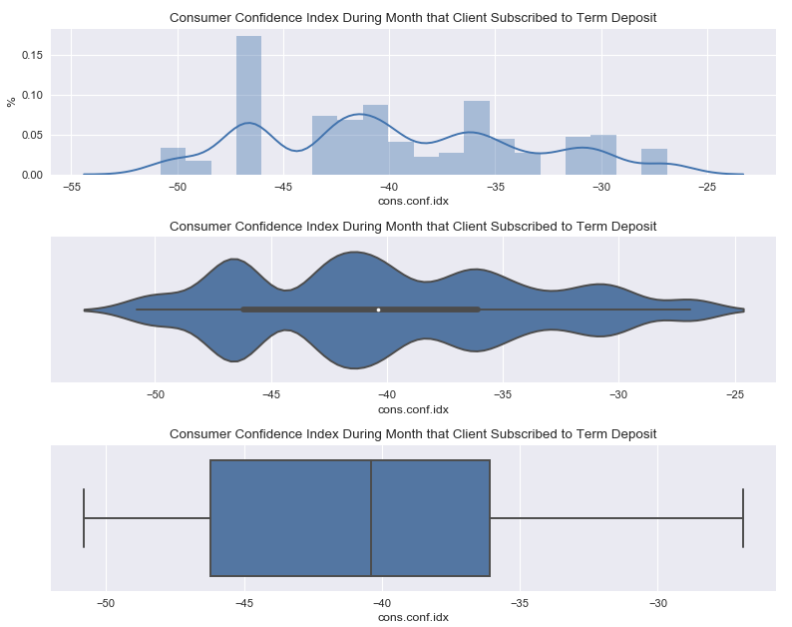
The general trend visible from the employee variation rate visualizations and table is that as employee variation increases, clients are less likely to be investing money in term deposits.

**Consumer Price Index**

****

****

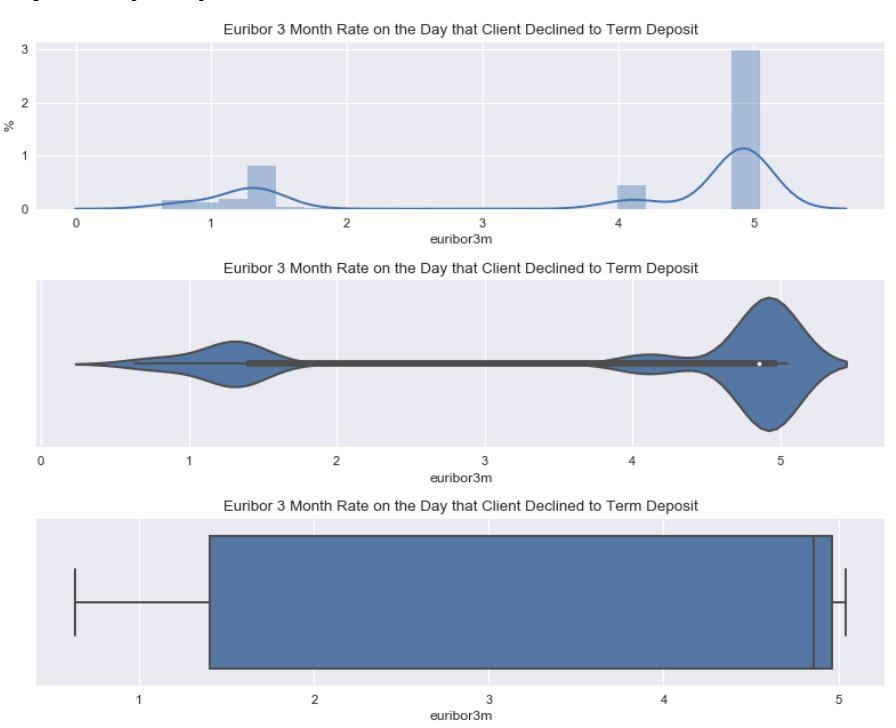
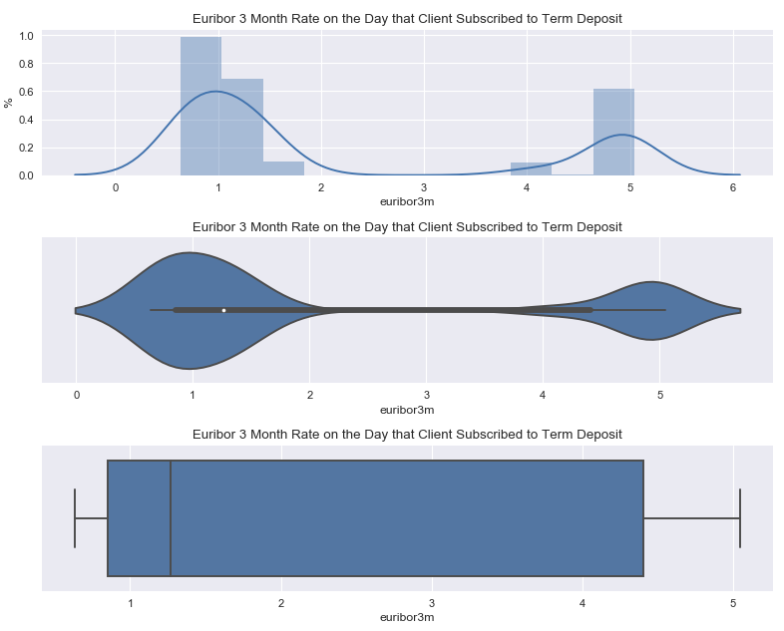
The distribution of clients that subscribed given different levels of consumer price index is distributed bi-modally with a slightly larger grouping of subscriptions at the 93.0 mode. Overall the average consumer price index is lower for those that subscribed than for those who did not. The distribution 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.

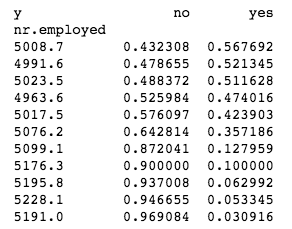
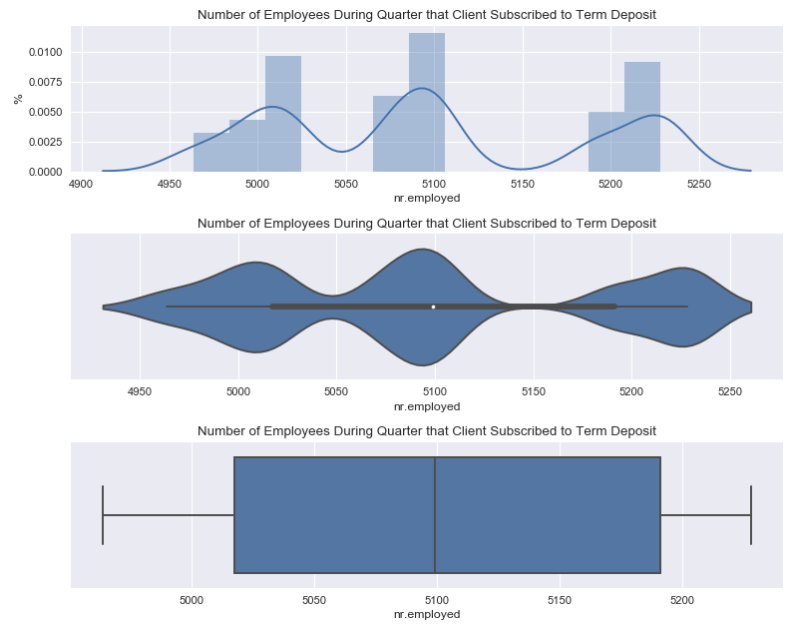
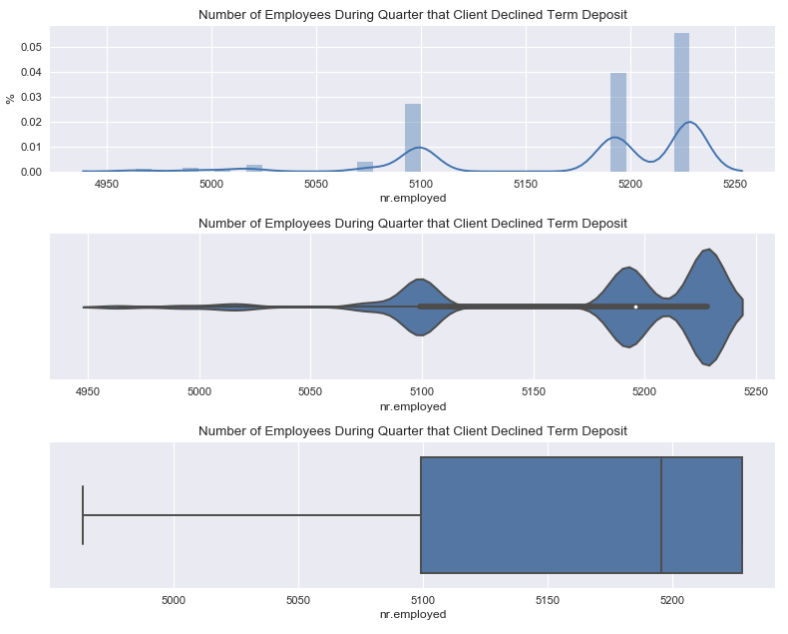
**Consumer Confidence Index**

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**

**Performance Metric**

As mentioned in the problem section, the primary performance indicator of the model will be the AUC of the ROC curve, because it will represent the bank’s goals of maximizing sales with less regard for the possibility of bothering the customer or spending more on telemarketing expenses. In addition to the AUC of the ROC curve, I will also look at accuracy as a general baseline of performance and because Sci-kit learn’s models and parameter tuning library are built for tuning for accuracy. Finally, I will also include some analysis of the AUC of the PR curve to address the hypothetical but real world problem of costs associated with bothering clients and costs associated with telemarketing campaigns.

**Feature Engineering and Selection**

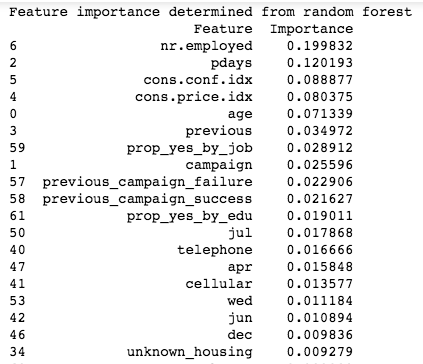
I elected to engineer two types of news features: custom binned features and proportions of positive responses to the target by category.

For the custom binning features I re-binned the following features: age, job, education, month of last contact, number of contacts performed during current campaign, number of days passed since last contact, and number of contacts performed before this campaign. The method I used to re-bin was to compute the proportion of positive responses by category and then group together features that displayed similar positive proportions of positive responses. Some features, like week day of last contact, did not have enough variation with regard to proportions of positive response rate to be worth re-binning.

For the proportion or positive responses by age, education, marital status and job I waited to add this feature until after the data was split into train and test sets to prevent data leakage. I computed the proportion of positive responses to the target variable for each group within the feature i.e. age, job type, etc. and then merged these proportions to the test data set by appropriate category group.

After creating the above mentioned features, I conducted correlation analysis by plotting the heat maps and identifying any two features that had above a threshold of .75 coefficient of correlation. If two variables did display a coefficient of correlation above the threshold I then ran a random forest classification model and created a feature importance table to determine which of the two highly correlated features was more important. Finally, I would drop the lesser important feature and iterate the process until there were no highly correlated features remaining.

With regard to feature selection, all twenty-two features were significantly important. This was evident before starting the project, when I read the white paper that this project is based off. The original study started with 155 features and used a combination of domain expertise and forward selection methods to whittle the feature set down to 22.



**Feature** **Importance**

For sensitivity analysis I created a feature importance table from my best Random Forest Classifier which is displayed to the right. Immediately evident is the fact that of the top five features, three are social/economic indicators. The most important of these, the number of people currently employed (in the Portuguese economy) has a coefficient in the logistic regression model of -0.0109 and is statistically significant at 95% confidence. This suggests that as employment rises fewer people invest their money. Consumer price index has a coefficient in the logistic regression model of -.1329 and is also statistically significant. Meanwhile, the coefficient for consumer confidence index tracks oppositely to consumer price index and has a coefficient of 0.0173. Of the non-economic/social indicators, the next most important feature is the number of days that passed by after the client was last contacted from a previous campaign. This feature is statistically significant at a 95% confidence level and is negatively correlated with term subscriptions i.e. the fewer days that have passed since last contact the more likely the client is to accept the term deposit. Next most important feature is age. The coefficient for age is small at .0002 and is not statistically significant at a 95% confidence level in the logistic regression model. The third most important non-economic/social indicator is the previous number of contacts performed before this campaign. The greater the number of contacts, the less likely the client is to subscribe to the term deposit, but again in the logistic regression model, this feature was not statistically significant at a 95% confidence level. The next most important feature is an engineered feature that represents the proportion of people that subscribed to the term deposit by job. This suggests that type of job impacts clients banking needs. The greater the number of contacts during the campaign the less likely the client was to subscribe to the term deposit and this feature was statistically significant in the logistic regression model at 95% confidence level.

**Models overview**

I used the Sci-kit Learn library for developing models. I attempted building models with the following five different types of algorithms: Logistic Regressor(LR), Random Forest Classifier(RFC), Multi-layer Perceptron Classifier(MLPC), K-Nearest Neighbors Classifier(KNN), and a Support Vector Classifier (SVC). I elected to experiment with five different types of models, because each model has a set of positives and drawbacks associated with it. LR, RFC, and KNN provide a more explainable model, whereas MLPC and SVC tend to provide better performance with regard to the loss function, but lack on explainability. In addition, some of models specifically in the Sci-kit learn library have different built in attributes and methods which present another layer of advantages and weaknesses. In the following sections I will dive into the specifics of each model and conclude the model overview section with a comparison of the five models.

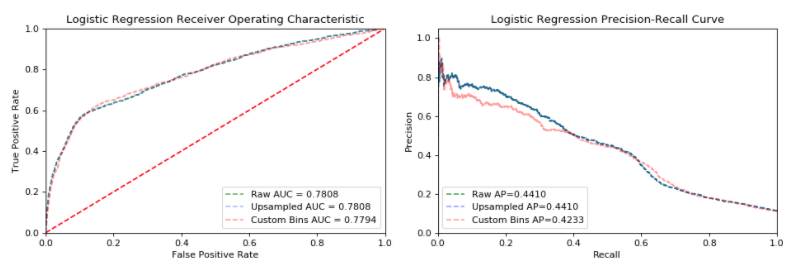
*Resampling and custom binning*

Due to the fact that the dataset was imbalanced with only 11% of samples being positive cases, I elected to resample the data using a random upsampling technique. This technique involved randomly selected with replacement samples from the minority class and adding them into the overall sample population when building my model. I build models both on the original set of data as well as on the resampled set. In addition to these two sets of data frames, I also included a third data set that consisted of the re-binned features from the feature engineering phase.

**Logistic Regression**

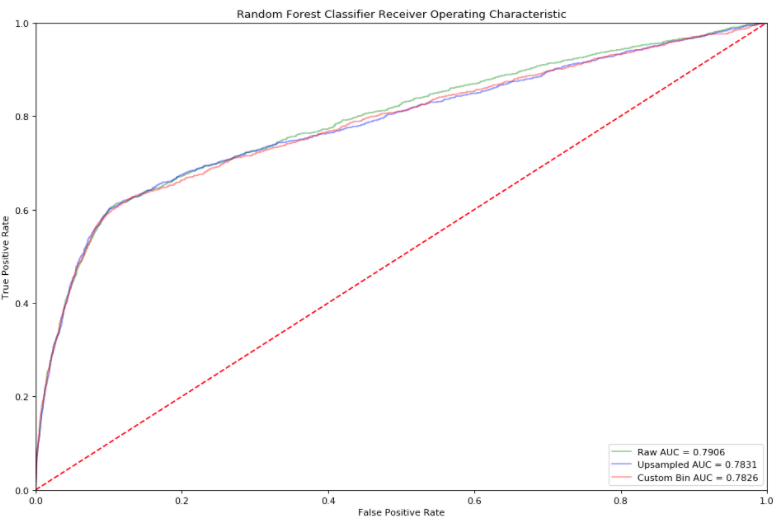
I started off modeling the data with a logistic regressor, because it is one of the most common classification algorithms and a good model to establish a baseline accuracy and AUC for. All three data frames tied at an AUC of .79. The raw and upsampled models tied at a PR AUC of .44 and a test accuracy of 90.19% both of which beat the custom bins model at a PR AUC of .42 and a test accuracy of 89.85%. For all three datasets I parameter tuned for the regularizing parameter C. The optimal C for raw data was 1, for the upsampled it was 500 and for the custom binned it was .01.

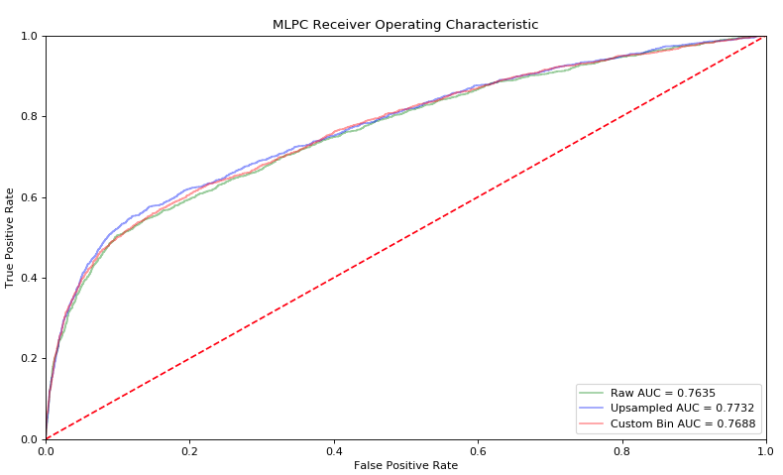
One unique characteristic of the logistic regression model is that it is possible to calculate and plot both the ROC and PR curves. In order to calculate the ROC curve it is necessary to calculate the probability estimates and to calculate the PR curve it is necessary to compute the confidence scores for the samples. In the Sci-kit Learn library the method Predict\_prob predicts the probability estimates of a classifier and the method Decision\_function predicts the confidence scores for samples. The logistic regression classifier has both of these methods whereas most classifiers only have one or the other. Another advantage to the Logistic Regression is that the summary of the predictor variables and statistical information about the model can be generated using the Scipy package. This allows for a high degree of model interpretability which in and of itself can help bank management optimize sales practices.

In looking at the below curves the ROC curve for al three datasets follow a similar path with only minor variation. The Raw and Upsampled plots follow a near identical path so it is difficult to decipher the two separate lines. There is more interesting variation in the PR curve with the Custom Bin data dropping below the other two curves from roughly .02 through .4, but then also having a short range where it is higher than the other two curves.

**Random Forest Classifier**

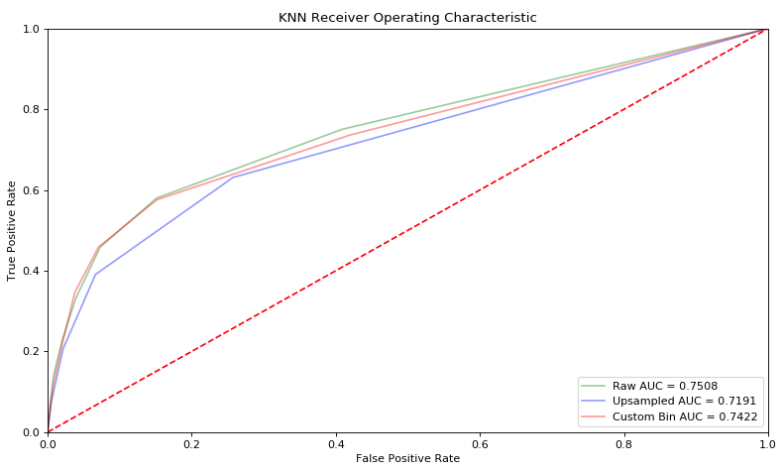
The second model I attempted was a Random Forest Classifier. For this model the raw dataset outperformed the other two by approximately .01 on the AUC. The optimal parameters from parameter tuning were 360 estimators, 4 minimum sample leaf, auto max features and a max depth of 15. The RFC in the Sci-kit library does not have the ability to calculate the PR curve easily and it is not the primary performance metric of this project so it is not included. All three lines follow a similar trajectory except from a false positive rate of between approximately 0.35 and .9 during which the raw data set outperforms by a slightly higher true positive rate. Random Forest Classifiers are robust against class imbalances so the fact that the raw data outperforms the upsampled data is not unexpected. Similarly to the Logistic Regression, the RFC also provides informational insight into the predictive features, but via the feature importance method. This feature importance method was analyzed previously, but in assessing this model it adds to the models viability in a business context, because it allows for improved explainability to bank management.



**Multi-Layer Perceptron Classifier**

For the Multi-Layer Perceptron Classifier the interesting result was that the raw dataset that performed highest for the RFC and tied for highest for LR performed worst in this model. All three models again follow a similar line trajectory with the upsampled data set having a noticeable upward deviation between 0.1 and 0.4. The upsampled data model slightly outperformed the custom binned data and had the highest AUC at 0.7732.

In terms of feature tuning I tuned the following parameters: max iterations, hidden layer sizes, batch size, and alpha. For the upsampled dataset the optimal parameters were 1750 iterations, 4 layers, a batch size of 100 and an alpha of .01. Interestingly, despite having the highest AUC the upsample dataset had the lowest test accuracy at 62.97% compared to the other two that were both near 90%.



**K-Nearest Neighbors**

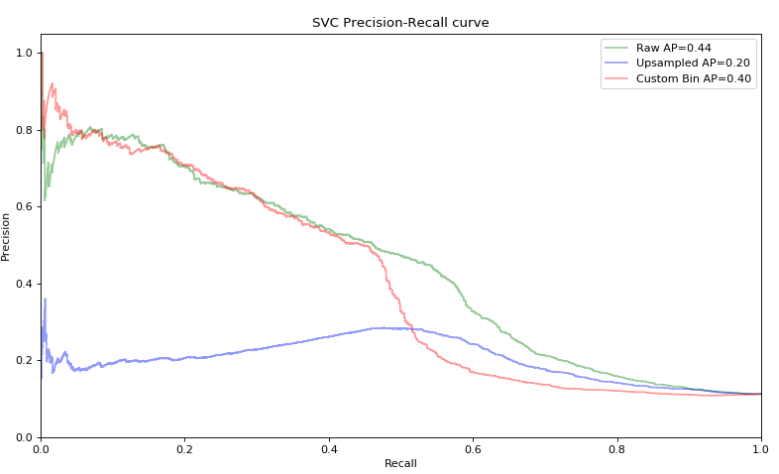
For the K-Nearest neighbors algorithm the unique characteristic from the previous models is that the AUC scores for the three datasets was more significantly different with the raw dataset placing first at an AUC of .7508. For parameter tuning I elected to tune for the number of neighbors and the leaf size. For the raw dataset the best model had a nearest neighbors of 8 and a leaf size of 20 and the high training score of 89.55%.

**Support Vector Classifier**

The support vector classifier is the only of the three that does not have the predict\_proba method and thus it is not easy to calculate and plot its ROC curve. This makes it impossible to effectively compare this model to the others.

For parameter tuning, I elected to tune gamma, the kernel coefficient and C, the penalty parameter. The data frame that performed the best was the raw dataset with a AUC PR of 0.44. The optimal gamma and C value for this raw dataset was 0.00001 and 1000 respectively. Its test accuracy was 90.00%.

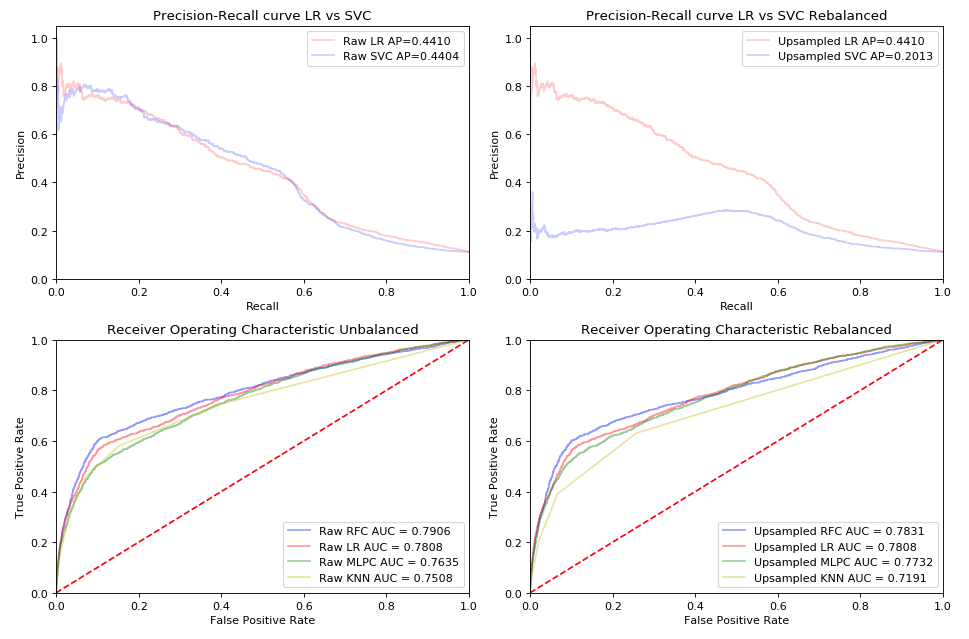
For this model in the beginning portion of the PR curve the custom bin data frame significantly outperforms the raw data frame, but starting at approximately a recall of .1 the to lines converge and follow a similar trajectory up until just before a recall of .5 at which point the precision of custom dramatically drops.

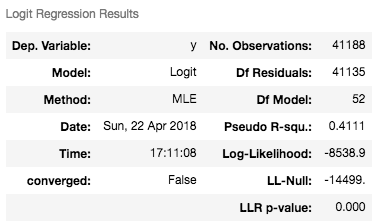


**Comparing the models and Recommendation**

In comparing all of the models, the raw RFC model performed the best with an AUC of 0.7906. Its superior AUC is due to its high true positive rate over a range of false positive rates between 0.0 and 0.4 over which it is clearly the best model. At around 0.4 the raw logistic regression model begins to converge on the line and then over take it briefly.

It is interesting that both LR and RFC which are models that have advantages over the other models in explainability as oppose to performance both outperformed MLPC and KNN. My suggested model for the bank managers would be the RFC for the raw dataset, because it provided the highest AUC and also information regarding feature importance that could be leverage to optimize sales.



**Appendix**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **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 |

