**Bank Marketing Machine Learning Project**

Jennie Hirsch

**Introduction**

Machine learning algorithms are increasingly being deployed to drive efficiency and growth across industries. They can yield a range of customer insights, help prevent fraud and improve profitability1. These results are particularly pronounced in the field of marketing where predictive modeling and data mining contribute to personalization and lead scoring for more efficient marketing campaigns.2

The data for this project comes from a direct phone marketing campaign administered by a Portuguese banking institution from 2008 to 2010.3 The goal of the project is to build the best model for predicting whether a client will buy a term deposit following the telemarketing campaign and recommend a strategy for more efficiently choosing clients to call.

A variety of machine learning techniques for classification are compared, including logistic regression, random forests, and SVM’s. We split the training data into training, validation and test sets so we can do hyperparameter optimization on the random forest and SVM models. For model validation we use the ROC curve and AUC. We also look at misclassification rate and cumulative gains curve. The tuned SVM had the best performance with highest AUC, but logistic regression came in close second and was much faster to implement. Using the final SVM model we can improve marketing efficiency by ranking new data and only calling those customers with a higher probability of subscribing to a term deposit. For example, calling the top 50% of clients is estimated to yield 77% of the subscriptions.

**Description of Data**

The dataset has 45,211 observations representing unique clients called in a bank marketing campaign between May 2008 and November 2010. There are 16 features plus one binary response representing success of failure of the campaign. A successful phone call results in the customer subscribing to a term deposit. The success rate was 11.7%. The feature names and descriptions are listed below.

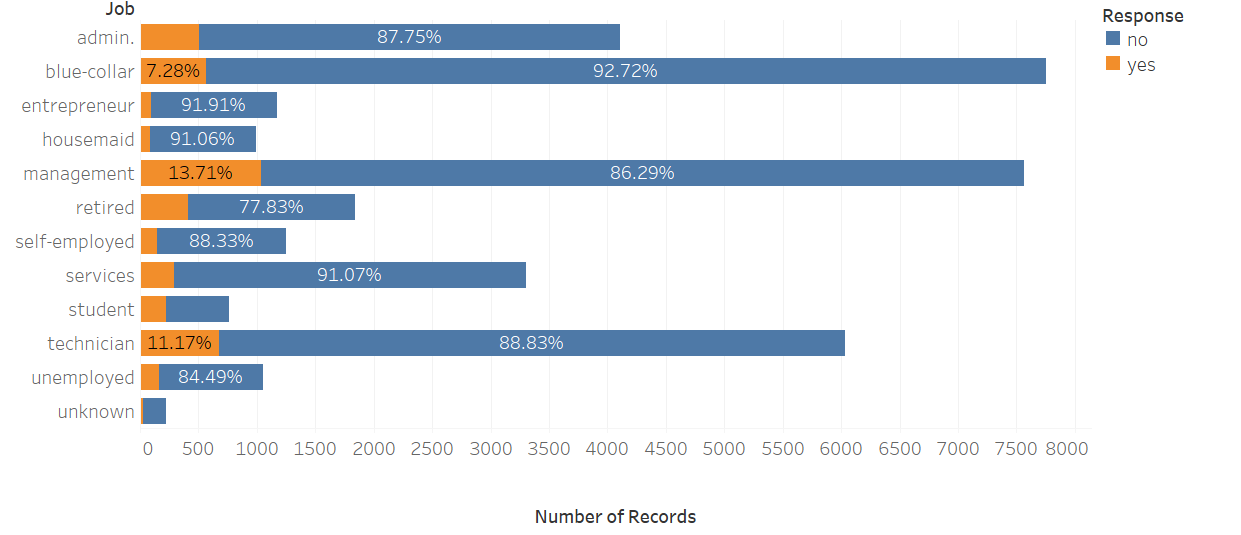
|  |  |
| --- | --- |
| **Feature** | **Description** |
| age | Numerical |
| job | One of 12 categories |
| marital | Single, married, divorced |
| education | Primary, secondary, tertiary, unknown |
| default | Whether or not the customer has ever defaulted on a loan |
| balance | Current bank account balance |
| housing | Whether or not the customer has a mortgage |
| loan | Whether or not the customer has a personal loan |
| contact | Whether they were contacted via cell phone or land line |
| day | Day of week |
| month | Month character |
| duration | Duration of last call of the campaign in seconds |
| campaign | Number of times the client was contacted in this campaign |
| pdays | number of days that passed by after the client was last contacted from a previous campaign |
| previous | number of contacts performed before this campaign and for this client |
| poutcome | outcome of the previous marketing campaign |

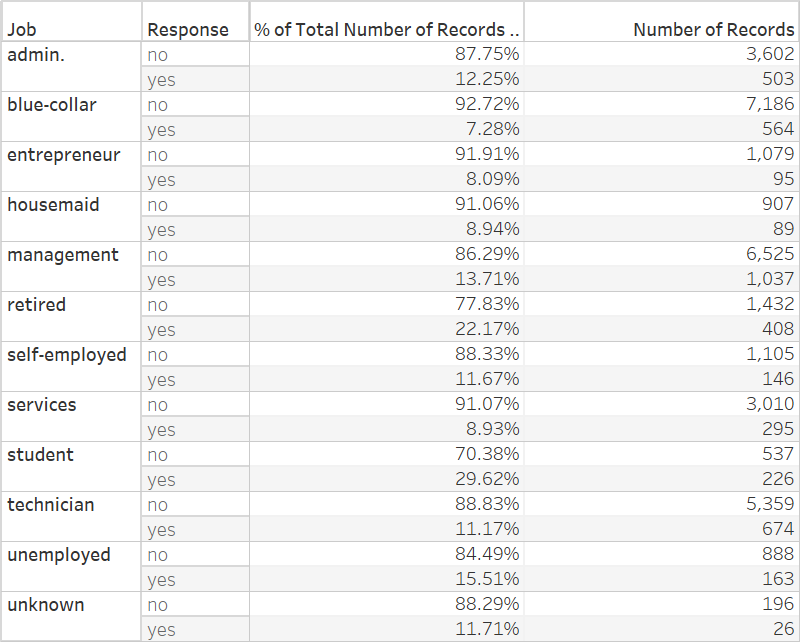
**Exploratory Data Analysis**

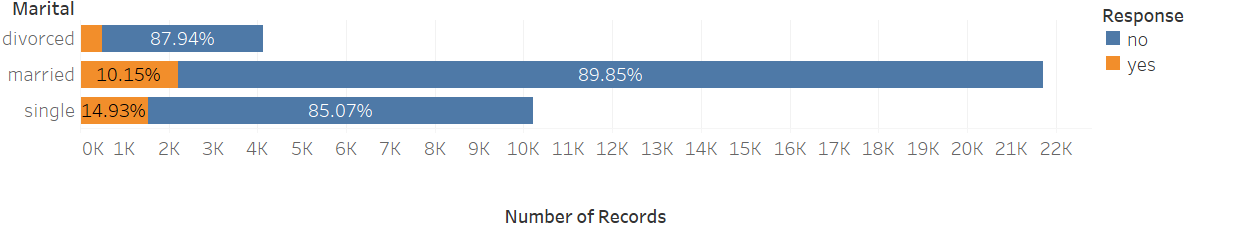
We split the data into train, validate and test sets using a random 80-10-10 split and perform exploratory data analysis on the training data.

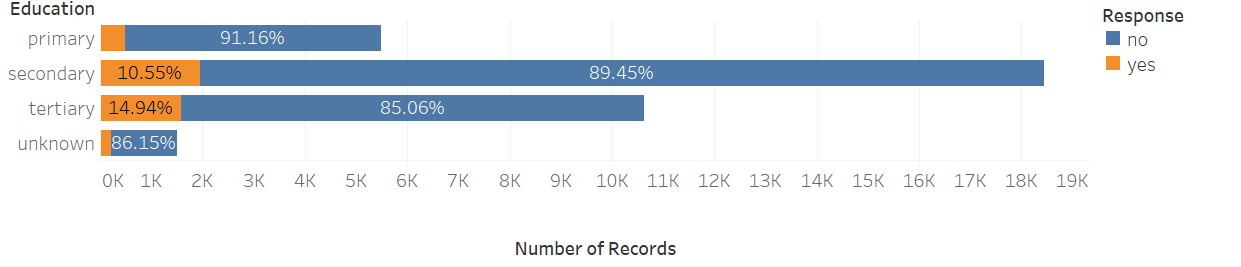
The goal of this analysis is improve efficiency of future marketing campaigns by predicting which customers are more likely to subscribe to a term deposit. This leads to the natural elimination of variables that cannot be known prior to calling like duration and campaign, as well as those that are not characteristics of the customer like month and weekday.

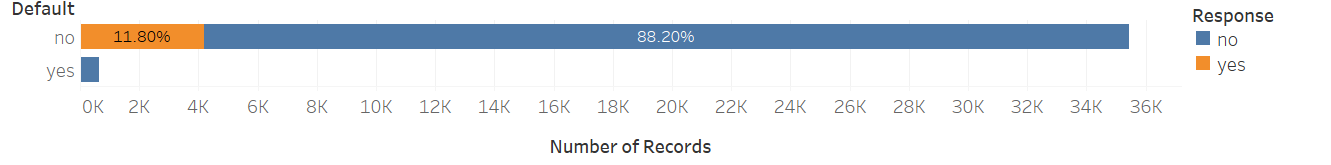
Next, we look at one variable at a time and try to understand which variables may have a larger impact on the response and if there are any variables we should eliminate from the analysis. Beginning with categorical variables, consider the following Tableau charts and tables showing the percent of successes and failures and number of records for each level of each variable.

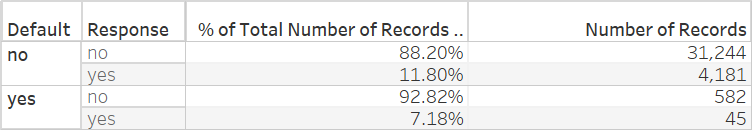


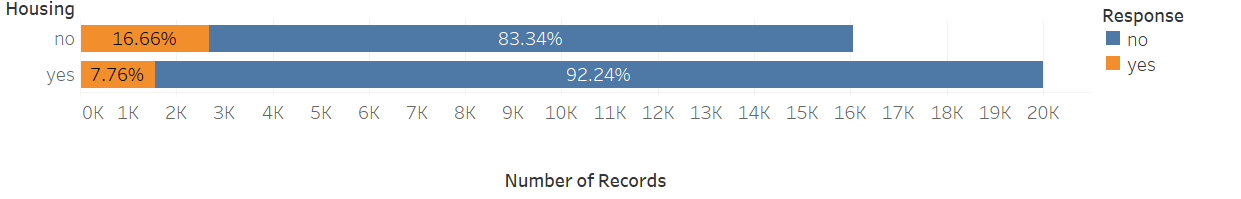


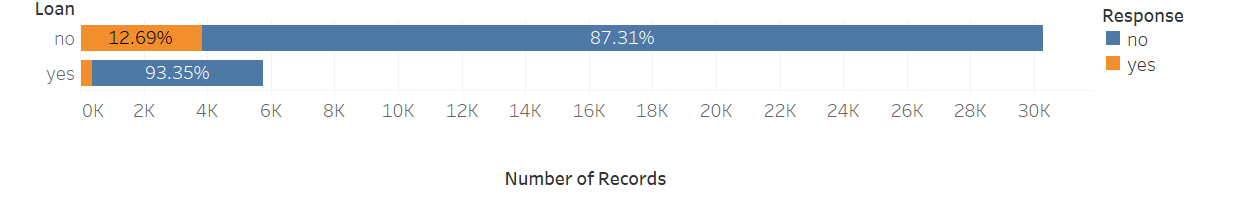


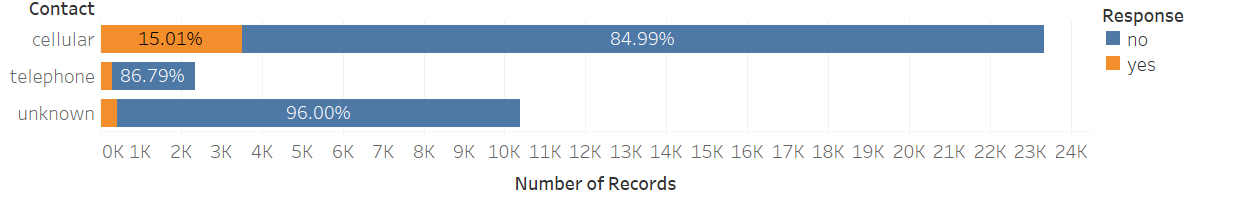


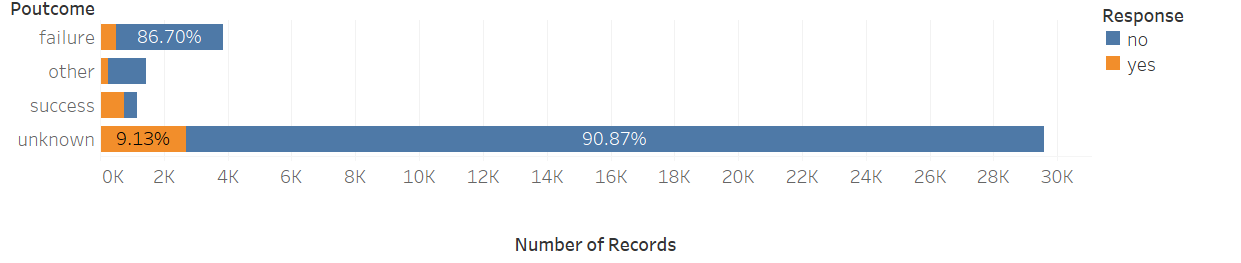


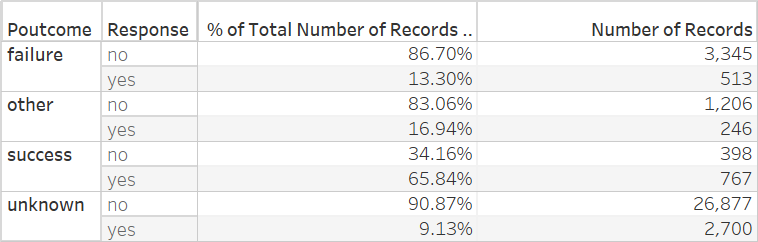






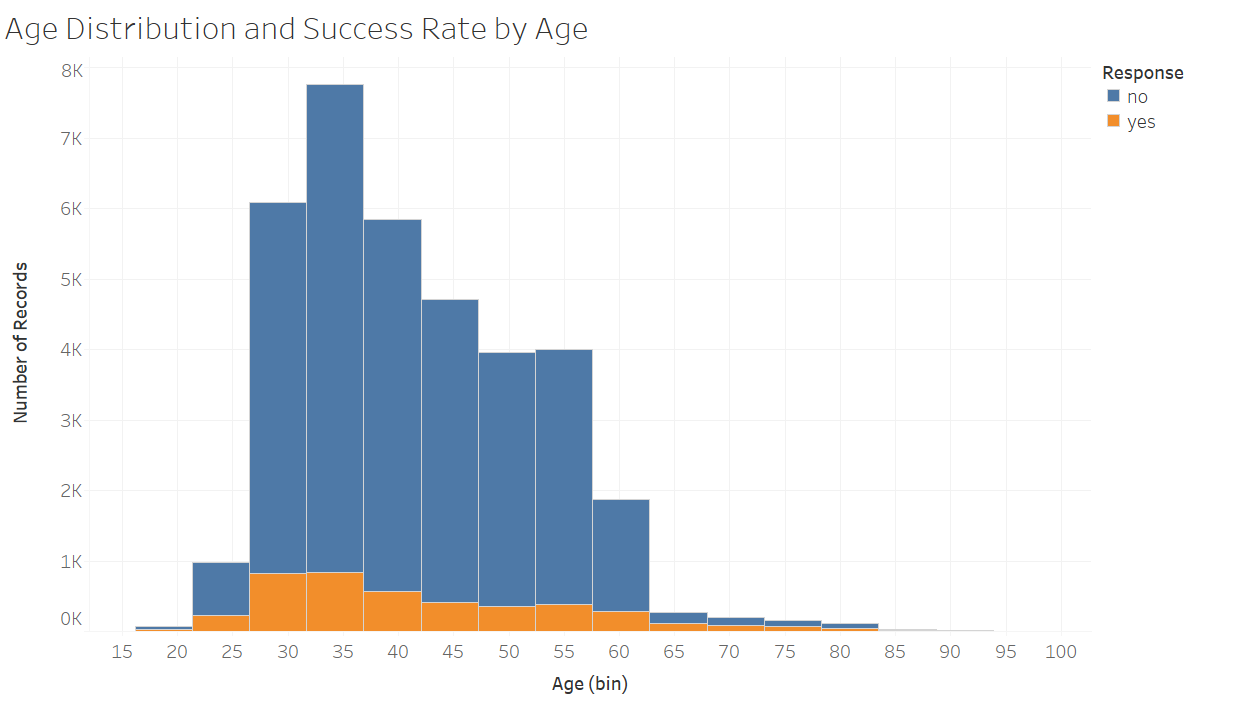


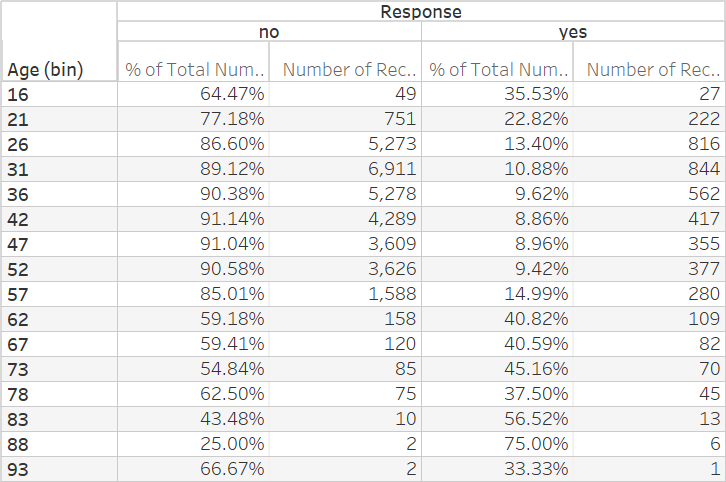


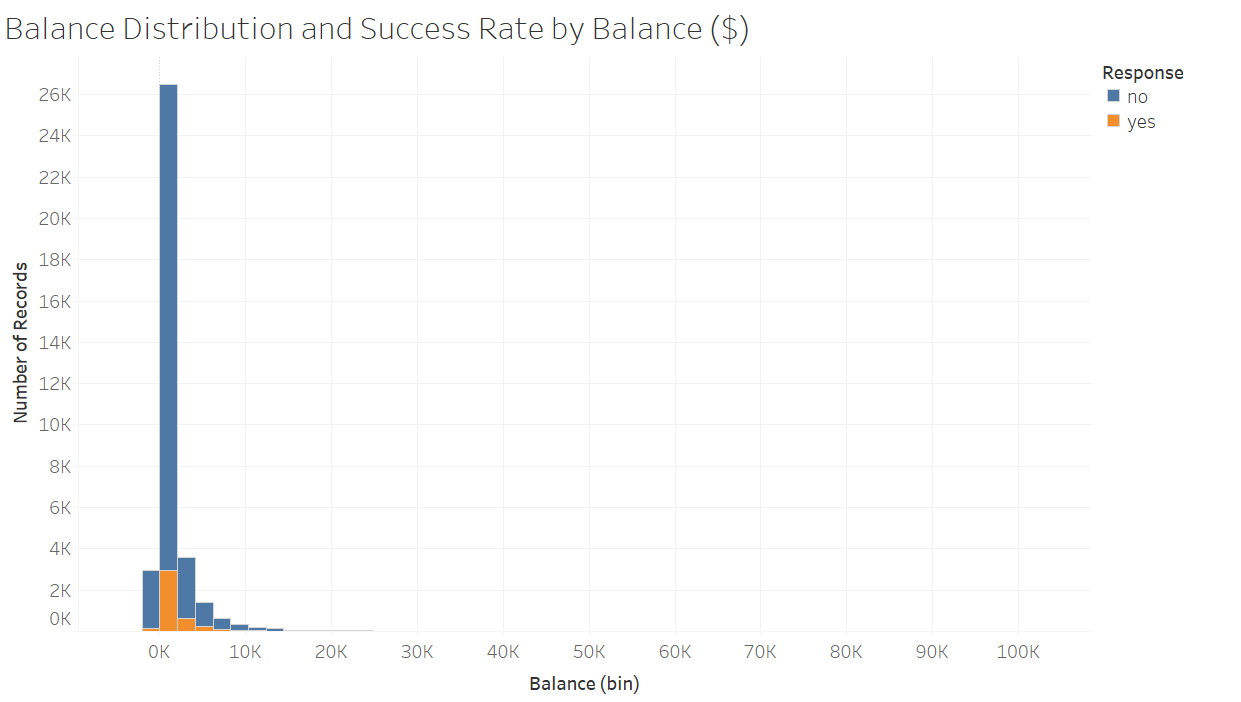


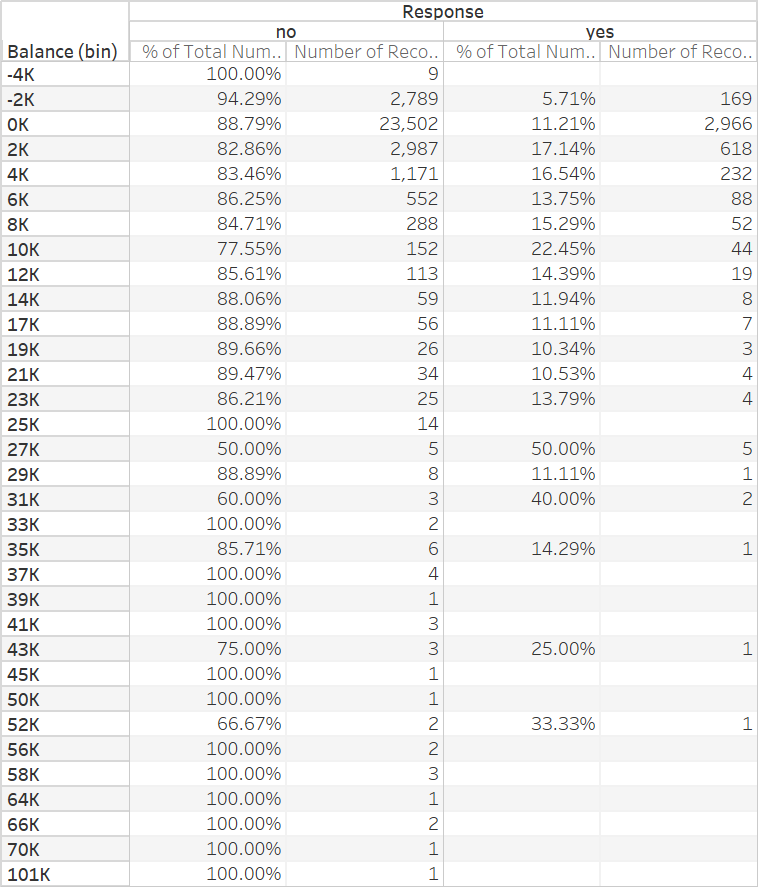
Many of the categorical variables appear important to the response at first look. Poutcome, job, and housing, are particularly influential. Those with a housing loan exhibit a 17% rate of subscribing to term deposits compared to the 12% average. Retirees and students have 22% and 30% success rates respectively. And customers who have subscribed to a term deposit in the past in response to telemarketing unsurprisingly show the highest success rate at 34%.

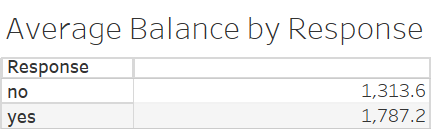
Now, let’s look at the numerical variables.

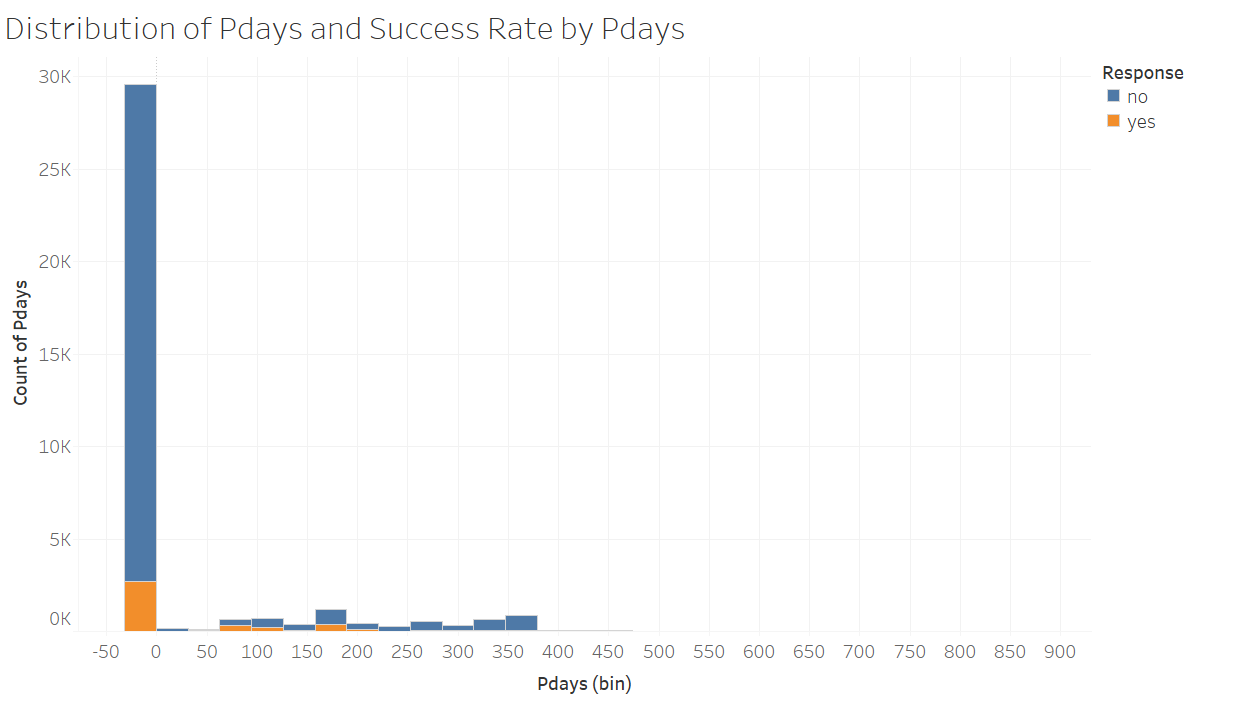


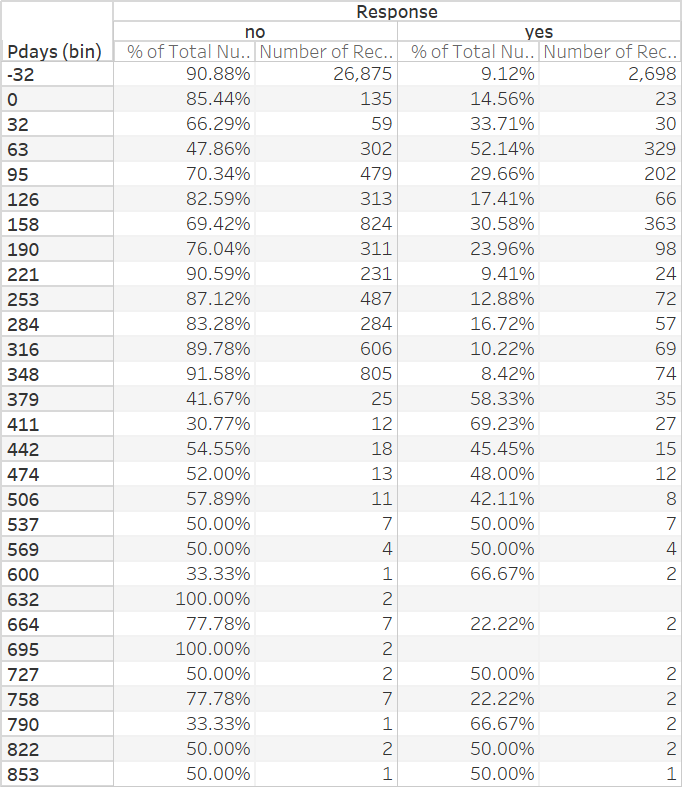








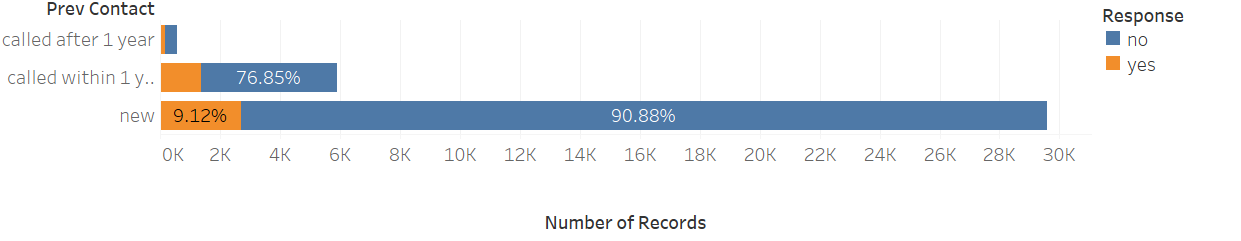


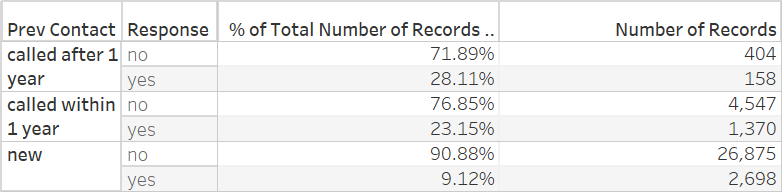


The age variable is an important factor in the response with both young and old customers more likely to subscribe to a term deposit than middle-aged. It does not appear that age is linearly related to the probability of success, so we consider including the square of age in the predictive models.

Subscribers to term deposits have $473 more in their account balances than non-subscribers on average. However, the success rate does fluctuate up and down for different account balances, so we try both second and third order balance terms in the predictive models.

Recall that the pdays variable is the number of days that passed by after the client was last contacted from a previous campaign. For simplicity and ease of implementation we convert this to a categorical variable called prev\_contact at taking on the values “new”, “called after 1 year”, “called within one year”. The best customers are those that have been contacted more than a year ago, followed by those that have been contacted within the past year. Finally, the variable “previous” has so much overlap with pdays that it was dropped for simplicity.





**Model Validation Metrics**

The business goal of this analysis is to improve the efficiency of future marketing campaigns. It is desired to rank a new set of customers from most likely to least likely to purchase a term deposit when called. We select our final model by calculating the area under the receiver operating characteristic (ROC) curve. We also look at the confusion matrix as well as the cumulative gains curve. For the KNN model we can use only metrics based on the confusion matrix since we are not ranking observations.

**Model Fitting and Evaluation**

**Logistic Model**

We fit a logistic regression model in R using the glm function with the following results. (See R code file for details)

Call:

glm(formula = formula1, family = binomial, data = train\_data)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.4163 -0.5044 -0.3873 -0.2644 3.1497

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.777e+00 2.747e-01 6.471 9.75e-11 \*\*\*

age -1.277e-01 1.014e-02 -12.590 < 2e-16 \*\*\*

I(age^2) 1.434e-03 1.106e-04 12.966 < 2e-16 \*\*\*

jobblue-collar -2.049e-01 7.179e-02 -2.854 0.004319 \*\*

jobentrepreneur -3.156e-01 1.248e-01 -2.528 0.011458 \*

jobhousemaid -3.619e-01 1.322e-01 -2.738 0.006177 \*\*

jobmanagement -1.605e-01 7.229e-02 -2.220 0.026433 \*

jobretired -7.872e-03 1.054e-01 -0.075 0.940443

jobself-employed -2.277e-01 1.103e-01 -2.065 0.038910 \*

jobservices -1.555e-01 8.217e-02 -1.892 0.058487 .

jobstudent 2.382e-01 1.088e-01 2.190 0.028537 \*

jobtechnician -1.782e-01 6.767e-02 -2.634 0.008446 \*\*

jobunemployed 7.383e-02 1.068e-01 0.691 0.489286

jobunknown -3.289e-01 2.356e-01 -1.396 0.162649

maritalmarried -2.206e-01 5.741e-02 -3.842 0.000122 \*\*\*

maritalsingle -2.970e-03 6.621e-02 -0.045 0.964224

educationsecondary 1.206e-01 6.346e-02 1.901 0.057344 .

educationtertiary 3.250e-01 7.369e-02 4.410 1.03e-05 \*\*\*

educationunknown 1.903e-01 1.020e-01 1.865 0.062243 .

defaultyes -5.340e-02 1.598e-01 -0.334 0.738259

balance 1.001e-04 1.240e-05 8.074 6.81e-16 \*\*\*

I(balance^2) -4.691e-09 8.330e-10 -5.631 1.79e-08 \*\*\*

I(balance^3) 3.782e-14 7.854e-15 4.815 1.47e-06 \*\*\*

housingyes -5.281e-01 3.911e-02 -13.501 < 2e-16 \*\*\*

loanyes -4.516e-01 5.848e-02 -7.723 1.14e-14 \*\*\*

poutcomeother 2.379e-01 8.724e-02 2.728 0.006381 \*\*

poutcomesuccess 2.279e+00 8.145e-02 27.979 < 2e-16 \*\*\*

poutcomeunknown 1.529e+00 1.018e+00 1.502 0.133183

prev\_contactcalled within 1 year -7.084e-01 1.113e-01 -6.365 1.95e-10 \*\*\*

prev\_contactnew -2.408e+00 1.020e+00 -2.360 0.018298 \*

contacttelephone -3.829e-01 7.380e-02 -5.188 2.13e-07 \*\*\*

contactunknown -9.644e-01 5.688e-02 -16.956 < 2e-16 \*\*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 26055 on 36051 degrees of freedom

Residual deviance: 22308 on 36020 degrees of freedom

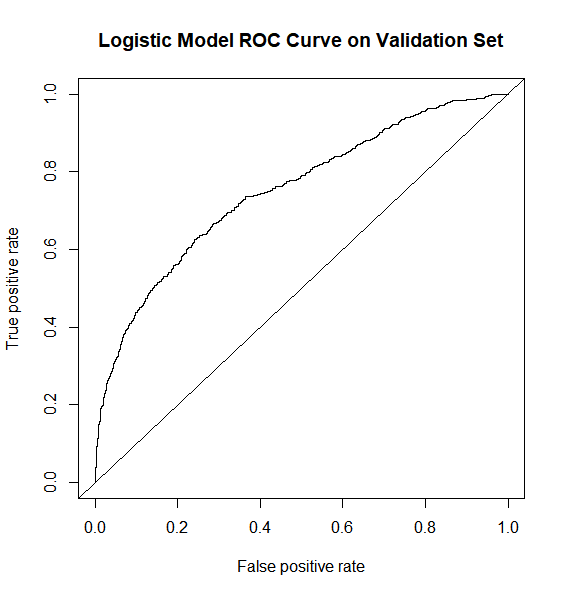
AIC: 22372

Number of Fisher Scoring iterations: 6

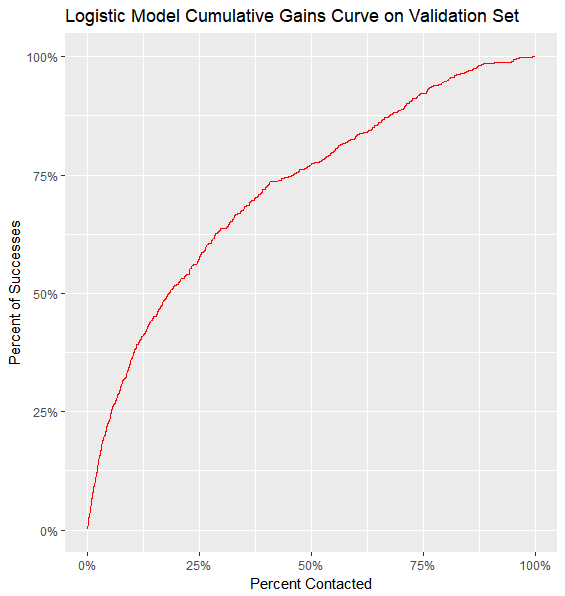
The logistic model finds many of the variables significant, especially the square of age, whether the customer is a blue-collar worker, account balance, whether the customer has a mortgage, and whether previous marketing campaigns have been successful. The confusion matrix, ROC curve, AUC, and cumulative gains curve are below.

|  |  |  |  |
| --- | --- | --- | --- |
| Logistic Model Confusion Matrix on Validation Set | | | |
|  | Actual | | |
| Predicted |  | No | Yes |
| No | 3,966 | 427 |
| Yes | 71 | 106 |

0.7507272



AUC = 0.7507272



This model looks good initially since its accuracy is 89%. However, only about 12% of the data points are successes so one could get 88% accuracy simply by predicting all failures. Luckily this model isn’t quite that simple. The ROC chart and area under the curve or AUC of 0.751 help quantify how much better this model does than a completely random one. The cumulative gains chart is also helpful in understanding how different calling strategies would lead do different results. For example we can call the best 50% of the customers and expect to get about 77% of the total subscriptions. This a definite improvement. Let’s see if we can do better with a different type of model.

**Random Forests**

We fit a random forest model in R using the randomForest package (See R code file for details). We first use the randomForest function with default parameters with the following results.

|  |  |  |  |
| --- | --- | --- | --- |
| Random Forest Model #1 Confusion Matrix on Validation Set | | | |
|  | Actual | | |
| Predicted |  | No | Yes |
| No | 3,946 | 423 |
| Yes | 91 | 110 |

MeanDecreaseGini

age 613.74184

I(age^2) 613.64577

job 490.21799

marital 153.47775

education 206.68277

default 16.43096

balance 872.96982

I(balance^2) 856.39530

I(balance^3) 876.93447

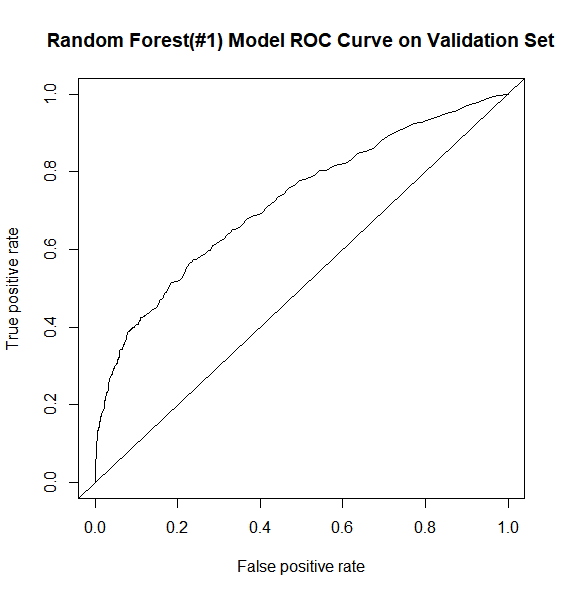
housing 149.30705

loan 72.92543

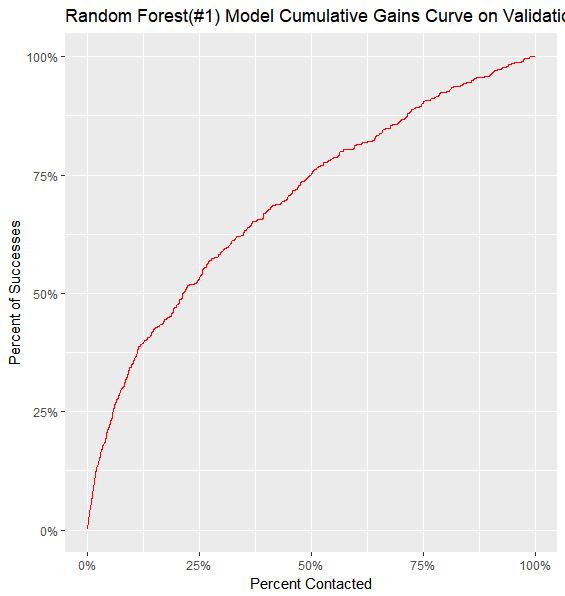
poutcome 587.70452

prev\_contact 164.61354

contact 162.23825



AUC = 0.72

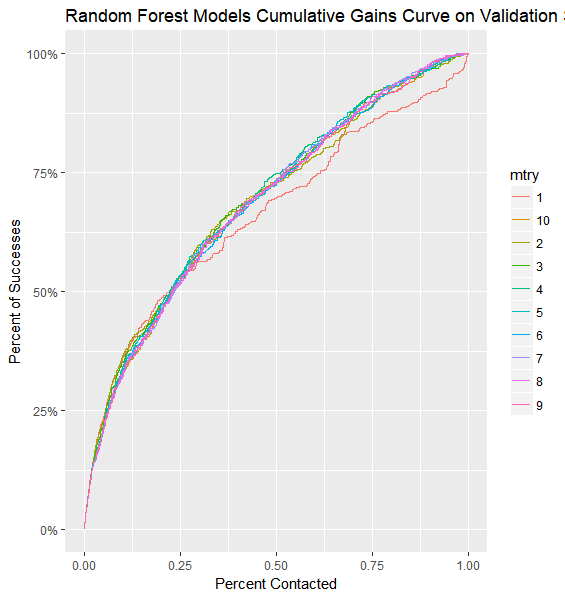


Like the logistic model this model finds balance, age, previous marketing campaign outcome, and job to be the most important variables.

The performance of this model quite as good as the logistic model. While the accuracy is 89%, the AUC is only 0.72 and if we were to rank the customers and call the top 50% we would only expect to get 75% of the subscriptions.

In order to improve the performance we try adjusting the number of variables that are used at each split of the trees in the random forest. We try values from 1 to 10 and we choose the parameter with the highest AUC on the validation set and also look at accuracy. The outputs from the code are below.

|  |  |  |
| --- | --- | --- |
| **Num of Var at Each Split** | **Accuracy** | **AUC** |
| 1 | 0.8879650 | 0.7013335 |
| 2 | 0.9072210 | 0.7195129 |
| **3** | **0.9076586** | **0.7210456** |
| 4 | 0.9078775 | 0.7203940 |
| 5 | 0.9089716 | 0.7193272 |
| 6 | 0.9089716 | 0.7126923 |
| 7 | 0.9087527 | 0.7162680 |
| 8 | 0.9107221 | 0.7144964 |
| 9 | 0.9098468 | 0.7131443 |
| 10 | 0.9096280 | 0.7123772 |

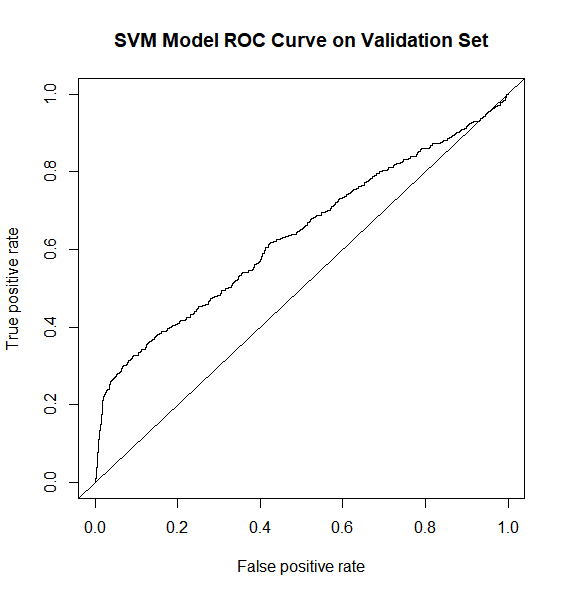


The random forest with the highest AUC uses 3 variables at each split. There are some values that produce slightly higher accuracy, but this is not quite as important. Since 3 is actually the default value, not much was gained from this exploration. We choose our first model as the best random forest but note that it does not perform as well as the logistic model. Perhaps the SVM will do better.

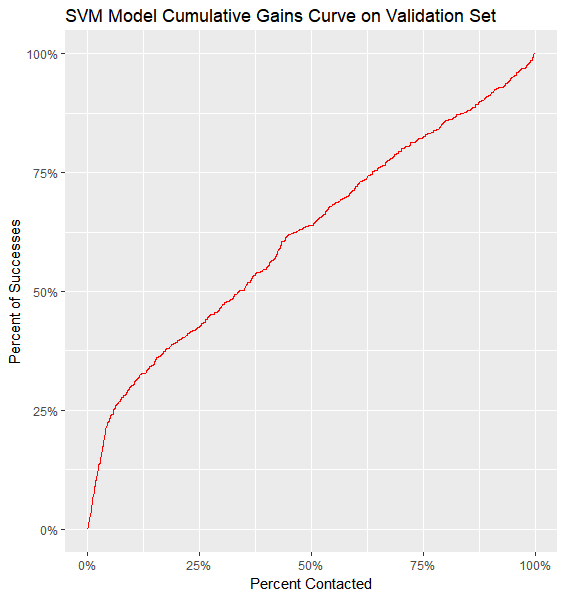
**Support Vector Machine (SVM)**

Using the e1071 library in R a SVM is fit to the training data using default parameters. The results are as follows.

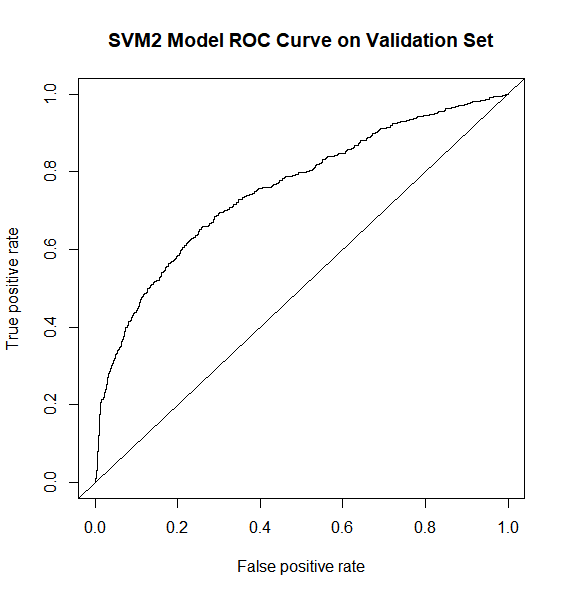
|  |  |  |  |
| --- | --- | --- | --- |
| SVM #1 Confusion Matrix on Validation Set | | | |
|  | Actual | | |
| Predicted |  | No | Yes |
| No | 3,963 | 421 |
| Yes | 74 | 112 |



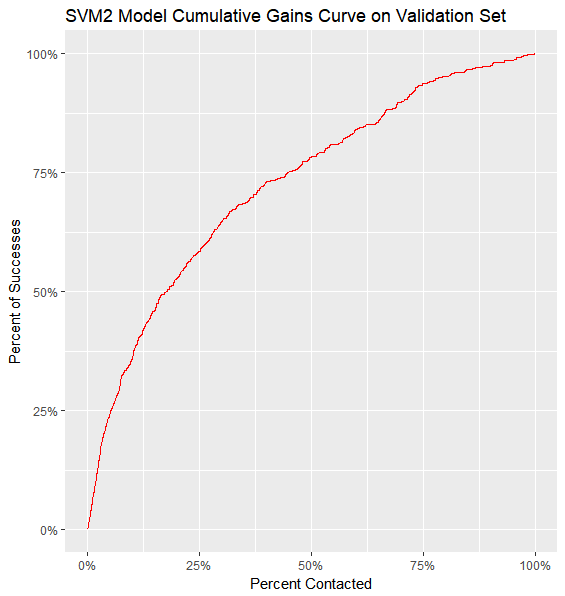
AUC = 0.6350463



The performance of this default model is not so good. One challenge with this dataset is that it is unbalanced -- there are not successes and failures in equal proportions. Since we are more interested in predicting successes and we only have 11.7% successes it would be ideal to adjust the model’s cost function to favor accurately predicting successes. This is done by adjusting the class.weights parameter in the svm function and re-fitting the model. Here are the results and we see lots of improvement.



AUC = 0.7564961



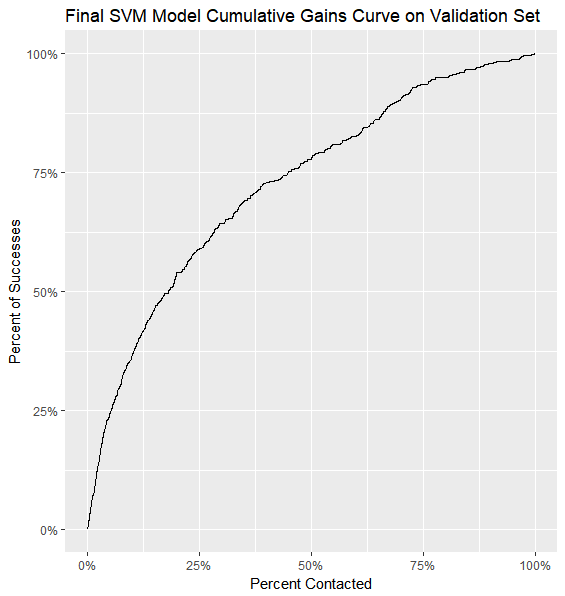
This new SVM model is the best model so far based on AUC and cumulative gains curve. Another common way to tune the SVM model is to adjust the cost and gamma parameters. Next, we fit SVMs over the grid where cost = 0.01,1,100,1000 and gamma = 0.1,1,10,100. Here are the results.

|  |  |  |
| --- | --- | --- |
| Cost | Gamma | AUC |
| 0.01 | 0.1 | 0.7233501 |
| 0.01 | 1 | 0.7186013 |
| 0.01 | 10 | 0.3234581 |
| 0.01 | 100 | 0.3456884 |
| **1** | **0.1** | **0.7540748** |
| 1 | 1 | 0.7279361 |
| 1 | 10 | 0.6766460 |
| 1 | 100 | 0.6531146 |
| 10 | 0.1 | 0.7254781 |
| 10 | 1 | 0.6560286 |
| 10 | 10 | 0.6354957 |
| 10 | 100 | 0.6288208 |
| 100 | 0.1 | 0.6970427 |
| 100 | 1 | 0.6403374 |
| 100 | 10 | 0.6170930 |
| 100 | 100 | 0.6228222 |

The best combination occurs when cost = 1 and gamma = 0.1, but this is not an improvement over the default parameters. Next, we try to hone in on the better part of the grid and fit SVMs over the grid where cost = 0.5,1,10,50 and gamma = 0.01,0.05,0.1,0.5. Here are the results.

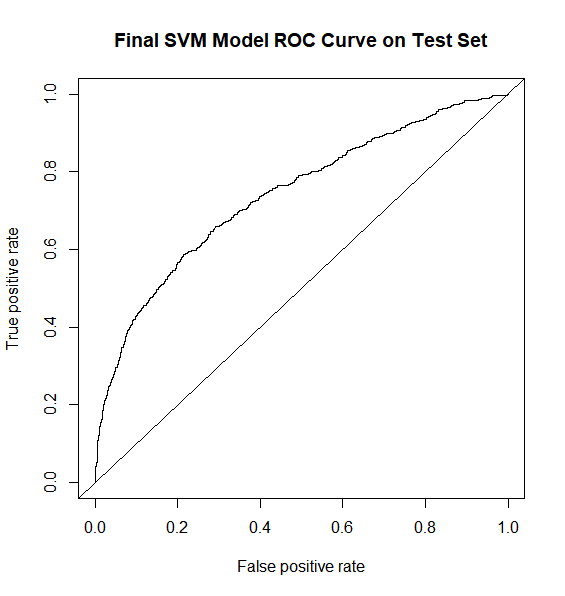
|  |  |  |
| --- | --- | --- |
| Cost | Gamma | AUC |
| 0.5 | 0.01 | 0.7447645 |
| 0.5 | 0.05 | 0.7554151 |
| 0.5 | 0.1 | 0.7539790 |
| 0.5 | 0.5 | 0.7473704 |
| 1 | 0.01 | 0.7500352 |
| 1 | 0.05 | 0.7550609 |
| 1 | 0.1 | 0.7541942 |
| 1 | 0.5 | 0.7418980 |
| **10** | **0.01** | **0.7566871** |
| 10 | 0.05 | 0.7536881 |
| 10 | 0.1 | 0.7481906 |
| 10 | 0.5 | 0.7062870 |
| 50 | 0.01 | 0.7520452 |
| 50 | 0.05 | 0.7473146 |
| 50 | 0.1 | 0.7330616 |
| 50 | 0.5 | 0.6784539 |

The new best model based on AUC is the SVM with cost = 10, and gamma = 0.01. If we call the top 50% of customers predicted by this model on the validation set, we expect to get 79% of the possible subscriptions.

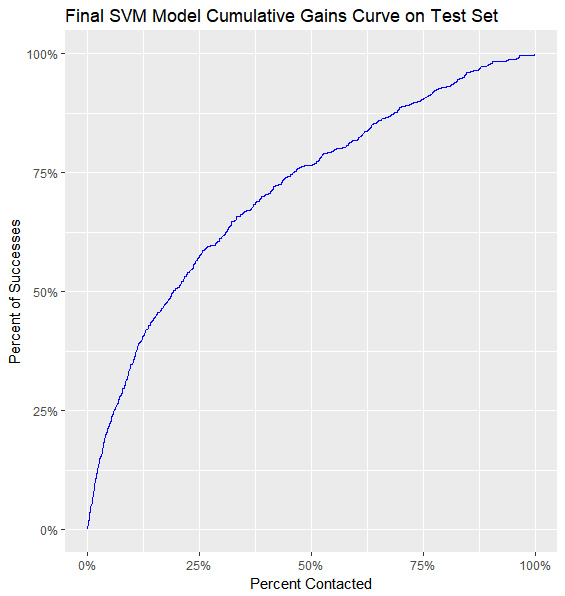


**Final Model Results and Conclusion**

As we have seen the SVM with cost = 10, gamma = 0.01, and class weights adjusted for the success rate in the training data. To arrive at this model, AUC was optimized on the validation set and the best model chosen. However, the validation set can’t be used to estimate the performance on new data. For this we use final 10% of the data in the test set. It is expected these results will not be quite as good as those on the validation set.



AUC = 0.7399157



The above graphs are the best estimates of how the final model will perform on new data. As expected the AUC of 0.74 is not quite as high as on the validation set. We can see from the cumulative gains curve that if we were to call the top 50% of the customers ranked by the model, then we would expect to get about 77% of the total subscriptions.

**Resources**

1. “Machine Learning – What it is and why it matters”, *SAS*, SAS Institute Inc., n.d., <https://www.sas.com/en_us/insights/analytics/machine-learning.html>
2. Columbus, Louis, “10 Ways Machine Learning Is Revolutionizing Marketing”, *Forbes,* Forbes, Feb 25, 2018, <https://www.forbes.com/sites/kpmg/2018/04/18/how-to-stay-ahead-of-disruptive-change-lessons-from-inside-an-innovation-lab/#4d61c07019a9>
3. [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014
4. S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS.