REPLACE WITH COVER PAGE

Table of Contents

SVM and Logistic Regression Modeling

Classification problems revolve around being able to predict what category a feature falls into within the dataset. The common tools for classification problems are:

- Percent correction classification (PCC): measures overall accuracy. Every error has the same weight.
- Confusion matrix: also measures accuracy but distinguished between errors, i.e false positives, false negatives and correct predictions.
- Area Under the ROC Curve (AUC – ROC): is one of the most widely used metrics for evaluation. Popular because it ranks the positive predictions higher than the negative. Also, ROC curve it is independent of the change in proportion of responders.
- Lift and Gain charts: both charts measure the effectiveness of a model by calculating the ratio between the results obtained with and without the predictive model. In other words, these metrics examine if using predictive models has any positive effects or not.

This project focused on the use of SVM as well as Logistic Regression Modeling using PCC, Confusion Matrix, and AUC-ROC.

Executive Summary

The Executive Summary of our model can be found in Figure 1 below. The conclusion is that the Logistic Regression Models before better than the Sub-vector-machine model.

The conclusion is that the Logistic Regression Model performed slightly better than the SVM in terms of accuracy. However, the SVM performed slightly better than the Logistic Model in terms of Specificity.

Figure 1: Model Summary

Metrics	Logistic Regression	Sub-vector-machine
Accuracy	82%	79%
Sensitivity	56%	45%
Specificity	90%	91%
ROC-AUC	81%	68%

While accuracy is an important metric to consider, sensitivity and specificity must be taken into account to ensure the model performs beyond just one class. The goal is to ensure the model refrains from performing poorly with the introduction of new data.

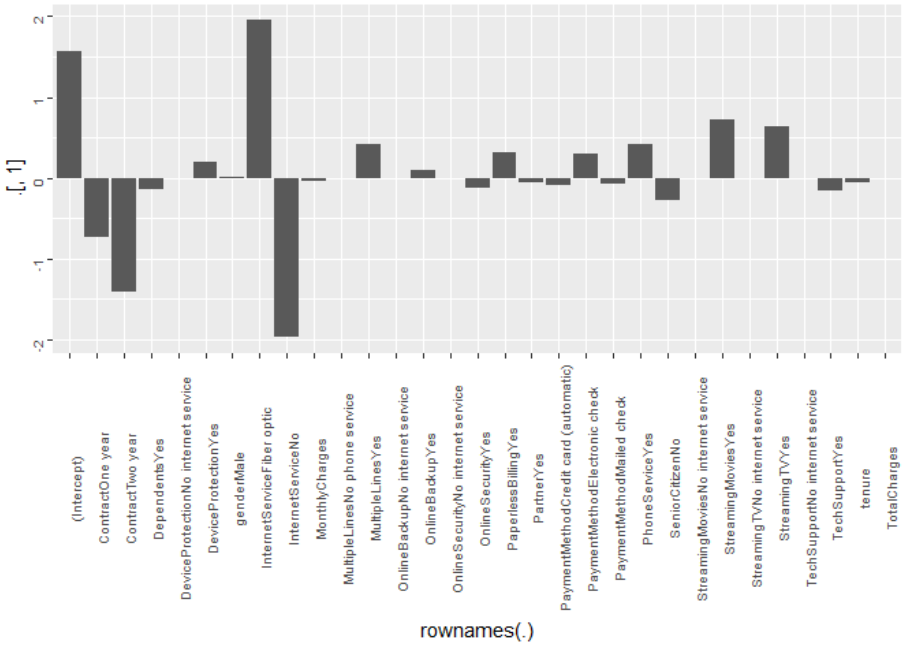
Details of the Models

While the details of the model performance and execution can be found in the Appendix, the various figures below will provide insight into the path followed by the Data Science team in exploring the data set and various models.

Full Logistic Regression Model

The initial goal is to determine the most influential factors in the data which end up being the type of Internet Service and the type of Contract. The Full Logistic Model unfortunately is better at determining loyal customers (or the probability when a customer won't leave).

Figure 2: Visualization of the Coefficients



The tools used automatically processes the 1-bit encoding of factor variables. This model however had some issues due to the fact that there is a high amount of correlation. For example, if a customer has no Internet service, they will always be absent of the ancillary services like Online Security or Online Backup.

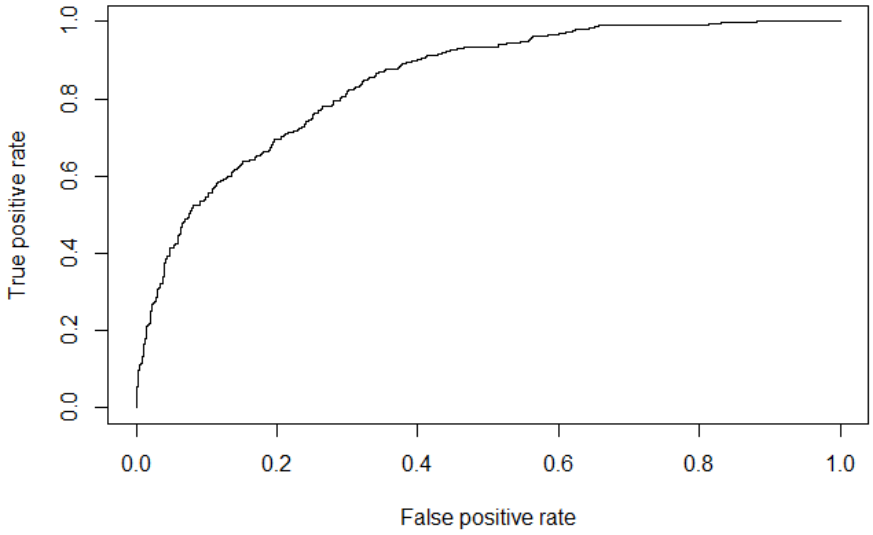
Despite the correlation issue, there are some parameters that emerge as more significant influencers in our model (all of these are deemed statistically significant):

- SeniorCitizen
- tenure
- MultipleLines
- InternetService
- Contract
- PaperlessBilling
- PaymentMethod
- TotalCharges

Total Charges, tenure, and contract variables are more significant than the other variables in due to their economic impact. The longer the customers stay with the phone company, the more they pay the phone company. The total charges will naturally remain higher for the loyal customers as their tenure in their contract status is also longer than those who churn.

In comparing the two models, the ROC curve tends to provide great insight into the accuracy and robustness of the model. Figure 3 provides insight into comparing the number of times the model predicts a yes or no comparison to the actual values. Accuracy is estimated at 82%.

Figure 3: ROC Curve for Full Logistic Regression Model



Along with visualizing the ROC Chart, a Figure 4 highlights a confusion matrix which will assist in determining the performance of the Full Model.

Figure 4: Confusion Matrix for the Full Logistic Model

Prediction	Yes	No
Yes	208	108
No	165	926

Due to the fact that the model is predominantly concerned with categorical predictors, it's not as obvious to identify potential correlation between them; as compared to continuous variables that allow the ability to calculate the Variance Inflation Factor [VIF].

Given the data model has over twenty (20) predictors, overfitting becomes a concern. This highlights the chance that the model performs well on the training data set (the one used to construct the model), but then performs poorly on new customers. Naturally, the goal is to accurately predict the customer's attrition probability.

There are tools used to help mitigate these concerns. The model should remain as simple as possible; leaving only the most influential predictors. In order to help refine the Full Logistic Model, the tools that can be used are:

- Feature Selection
- Regularized or Shrinkage
- Manually Adjusted

Stepwise Regression

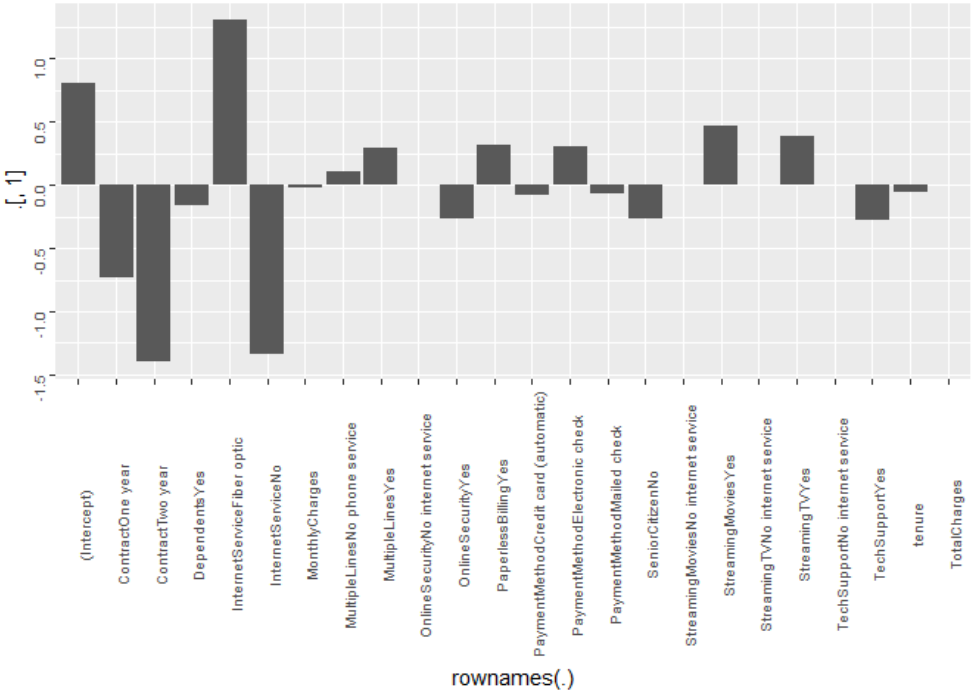
Stepwise regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure.

In each step, a variable is considered for addition to or subtraction from the set of explanatory variables based on some pre-specified criterion.

Figure 5 will outline the AIC based Stepwise Model recommendation of the following significant factors:

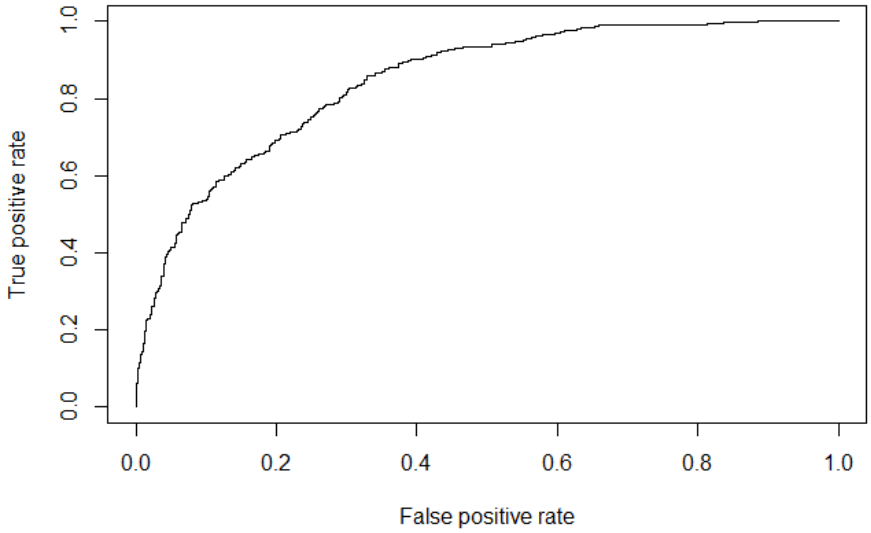
- SeniorCitizen
- tenure
- MultipleLines
- InternetService
- OnlineSecurity
- TechSupport
- StreamingTV
- StreamingMovies
- Contract
- PaperlessBilling
- PaymentMethod
- TotalCharges

Figure 5: Stepwise Model Coefficients



Using the Stepwise technique, the ROC curve can be calculated (displayed in Figure 6). The accuracy is still hovering around 82%.

Figure 6: Stepwise ROC Curve



The specificity (essentially predicting negative outcomes) is still around 90%. And the sensitivity, predicting positive outcomes, is slightly lower now at 58.98%

Figure 7: Confusion Matrix for the Stepwise Regression

Prediction	Yes	No
Yes	210	110
No	163	924

Lasso and Ridge Penalized Regression

Rather than removing predictors as indicated in the Stepwise model, the goal of Lasso/Ridge is to introduce a constraint in the training process that shrinks the estimated coefficients.

Ridge Regression introduces bias in order to reduce the model's variance. It then tries to minimize the sum of the Residual Sum-of-Squares (RSS) by shrinking the variables to approximately zero.

Lasso Regression is similar to Ridge, however it involves minimizing the sum of the absolute values of the coefficients, pushing the least significant ones closer to zero. Unlike Ridge, the Lasso Penalty will actually push them all the way to zero.

Both of these rely on a constant lambda which acts as a tuning parameter; affecting the size of the penalty. By affecting the size of the penalty, the coefficients's move closer to zero. A more detailed view can be seen in Figure 8 below.

Figure 8: Plot of Lasso Model Lambda

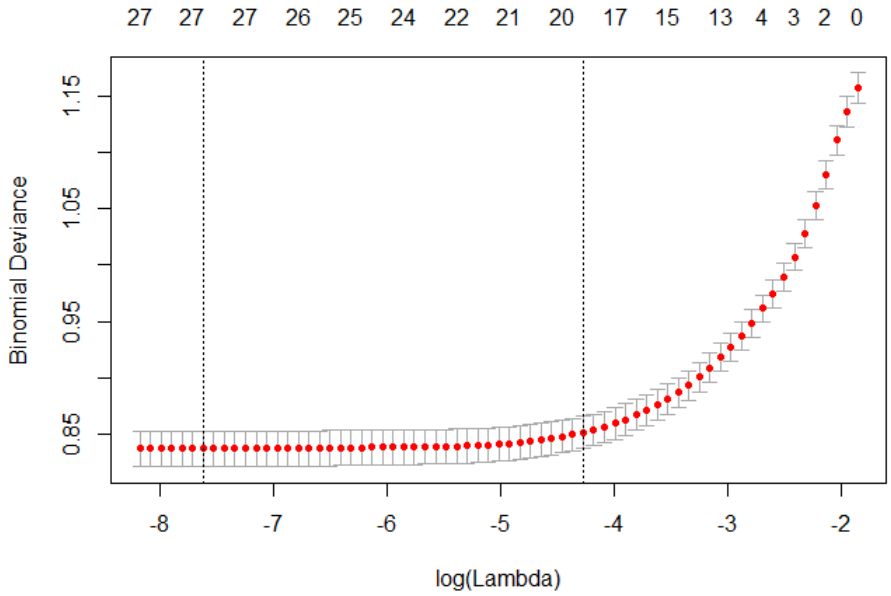
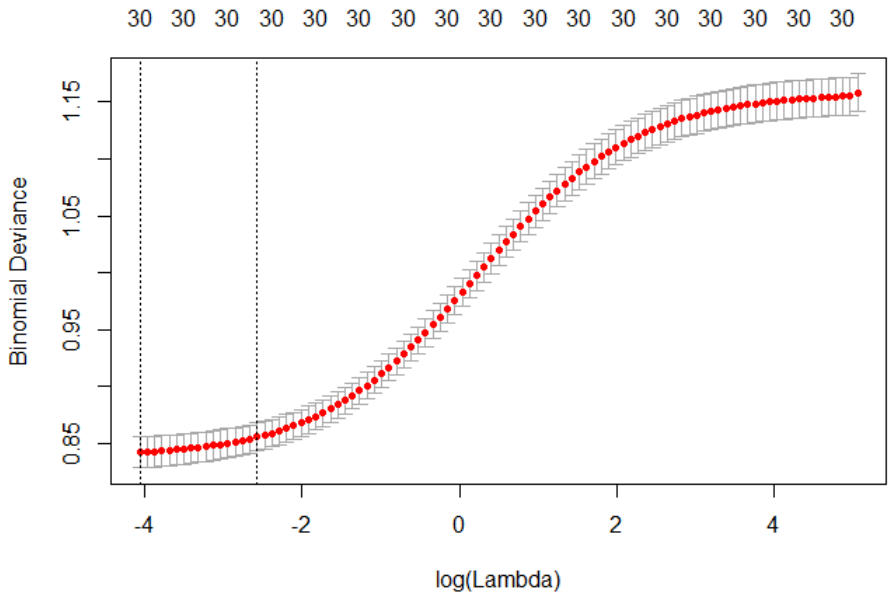
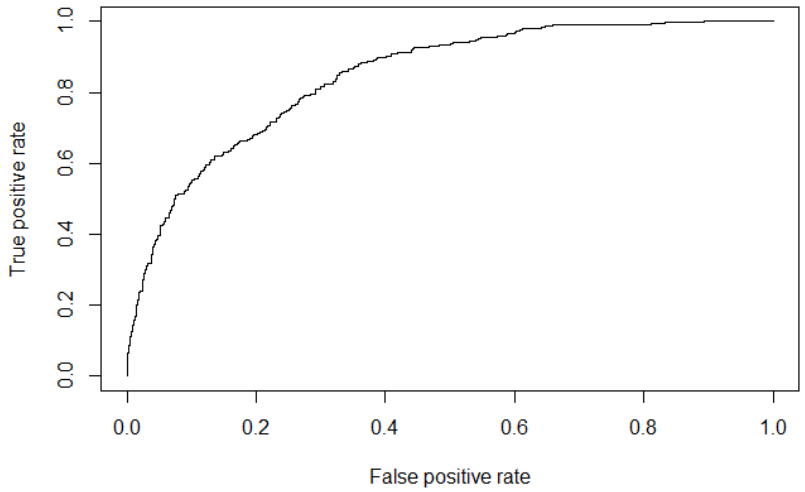


Figure 9: Plot of Ridge Model Lambda



In Figure 8 and 9, the first dotted line represents lambda with the smallest MSE and the second represents with an MSE within 1 standard-error of the minimum MSE. Balancing these comparisons with an ROC curve, Figure 10 shows that the model is still very closely aligned with the Full Regression Model.

Figure 10: ROC Curve for Lasso Model with Optimized Lambda



Comparing the number of times the model predicts an accurate result, either "Yes" or "No" in comparison to the actual values, we can see we have a slightly lower accuracy of approximately 82%.

Figure 11: Confusion Matrix for the Lasso Model

Prediction	Yes	No
Yes	207	108
No	166	926

Support-vector-machine Model

The Support-vector-machine model is designed for the supervised learning environment as in the Logistic Regression Model. The SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier.

An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall.

The model will employ an automated hyper parameter list generation to make C values candidates. It will then compare accuracy among different SVM kernel methods:

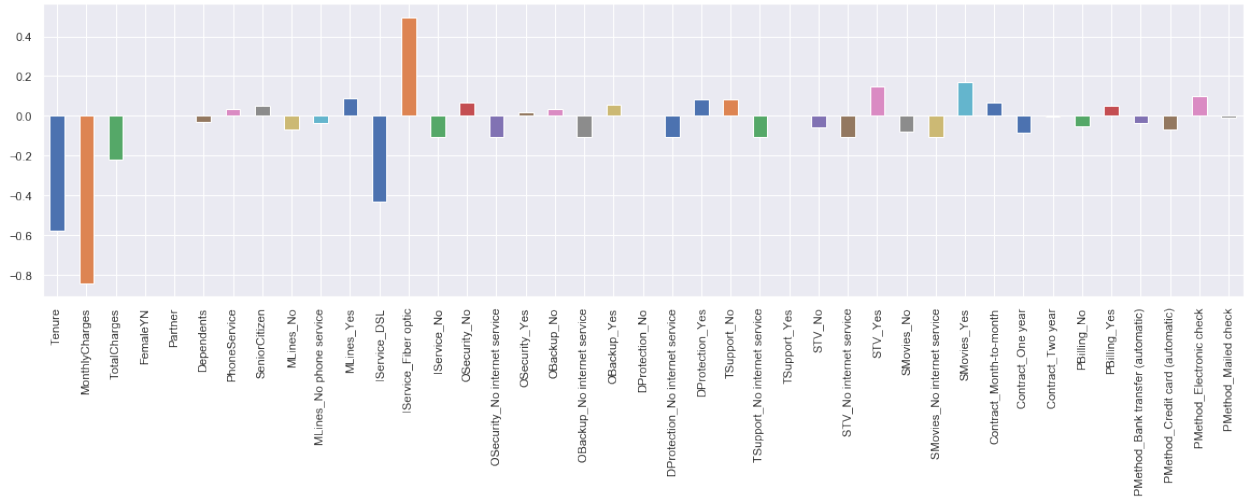
- Linear
- Polynomial
- Gaussian (RBF)
- Sigmoid

The automated model generated the sigmoid kernel with a suggested accuracy score and weights:

- Accuracy Score: 0.802484
- Using: "C" with grid result of 0.04642 with Sigmoid Kernel

Forcing a Linear Kernel results in the parameter list generation for weighted values in Figure 12 below.

Figure 12: Weighted Values from SVM-Linear Kernel Training



In order to avoid "snooping", the train data was chosen to be scaled to keep the test data intact. The automated model generated the sigmoid kernel with a suggested accuracy score and weights:

- Accuracy Score: 0.804969
- Using: "C" with grid result of 0.59948 with RBF Kernel

The training loop took 14 minutes (828 seconds). Given that the Logistic Regression model only took 38 seconds a conclusion can be made that SVM is definitely not as fast as the traditional Logistic Regression Model.

Advantages of the Models

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Interpreting Features

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliq

Support Vectors Insights

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliq

SVM and Logistic Regression Modeling

• [50 points] Create a logistic regression model and a support vector machine model for the

classification task involved with your dataset. Assess how well each model performs (use 80/20 training/testing split for your data). Adjust parameters of the models to make them more accurate. If your dataset size requires the use of stochastic gradient descent, then linear kernel only is fine to use.

• [10 points] Discuss the advantages of each model for each classification task. Does one type of model offer superior performance over another in terms of prediction accuracy? In terms of training time or efficiency? Explain in detail.

• [30 points] Use the weights from logistic regression to interpret the importance of different features for each classification task. Explain your interpretation in detail. Why do you think some variables are more important?

- **[10 points] Look at the chosen support vectors for the classification task. Do these provide any insight into the data? Explain.**