Management Summary For

MGT-415: Problem Set 3

Do Carmo Maria
Fan Xi
Genc Murat
Nyambuu Lkham
Trichilo Giulio

April 28, 2019

1. Introduction, Prediction Methodology and Feature Selection

In this report we outline our findings regarding the task of predicting churn, the dependent variable, on a set of 20 regressors in the Telco Churn dataset. For additional information regarding these regressors please refer to Report 1. We initially convert all categorical variables to numerical for computational purposes (as YES/NO and multiclass variables are treated as strings in the initial import). Binary variables essentially become dummy variables and multiclass variables retain integer values for each of their classes. The process of 'dummification' of multiclass regressors was not performed as to avoid the 'dummy variable trap' referring to multicollinearity arising from the procedure. We also perform an initial test of correlation between all variables in order gain a very basic overview of possible regressor interdependency (this is later formalized with Variance Inflation Factors).

Prediction Methodology

The task of prediction is considered both from a regression and a classification perspective, with particular regard to OLS and Logistic Regression. In order to formalize the prediction procedure, one should note that when performing OLS with a binary output variable (Y is churn or no churn) the OLS model becomes a Linear Probability Model. Therefore, in this particular setting we have that:

$$\begin{aligned} Y &= X\beta + \varepsilon \\ E[Y|X] &= E[X\beta + \varepsilon|X] \\ E[Y|X] &= X\beta \end{aligned}$$

Where X is the design matrix of regressors and β is the vector of OLS estimates. the last line follows from the Exogeneity assumption of OLS: $E[\varepsilon|X] = 0$. We assume this is validated as otherwise this would require an IV approach for which identification of instruments would likely prove to be time consuming with respect to the quality of the results to obtained, when assuming this holds. When Y is binary:

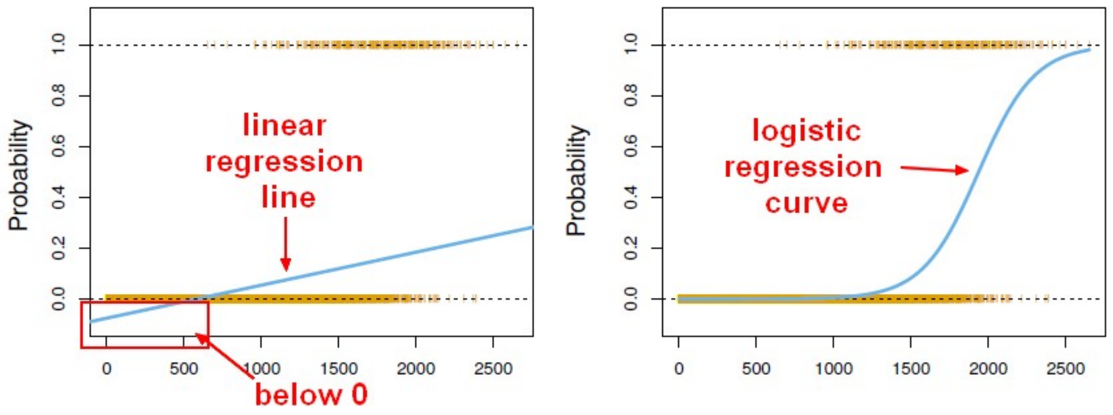
$$\begin{aligned} E[Y] &= 1 \times P(Y = 1) + 0 \times P(Y = 0) \\ E[Y|X = x] &= P(Y = 1|X = x) = X\beta \end{aligned}$$

Therefore predicted values $\hat{Y} = X\beta$ corresponds to the predicted probability of the output being 1, ie: a customer churns. As far as logistic regression is concerned, the logic is the same, except for that the linear function $X\beta$ is replaced with the logit function:

$$E[Y|X = x] = P(Y = 1|X = x) = \sigma(X\beta) = \frac{e^{X\beta}}{1 + e^{X\beta}}$$

This can be visualized in the representative image (in only one dimension) below:

As can be seen, the range of the regression line $X\beta \in \mathbb{R}^n$ can be outside of $[0, 1]^n$, which motivates the use for a logistic link function. In any case, we do compare linear regression and find that it is indeed weaker than logistic regression as well as other more complex classifiers.



Feature Selection

Feature selection can be decomposed into two main steps: removal of multicollinearity from the full design matrix, and then removal of statistically non-significant regressors from the selected model, then the resulting design matrix is split into train and test sets, and predictive performance is evaluated. Techniques such as Recursive Feature Elimination (RFE) could perhaps give us a better set of regressors, but at the cost of interpretability.

2. Regression

With regression (the linear probability model) we initially try to see whether there are significant differences in the quality of the fit when using X as: 1. The full dataset, 2. Only numerical variables, 3. Only categorical variables. From this we draw two initial conclusions: distribution of residuals is Bimodal, this per se does not violate the Gauss Markov theorem, we are however better off using heteroskedasticty robust covariance; it is interesting to note that this bimodality might be an indicator of different behavior between churn and non churn customers. As far as R^2 is concerned, all fits are roughly the same, therefore we have no reason not to use the full design matrix for what follows.

We then calculate the Variance Inflation Factors for all the regressors in X, and notice that the only abnormal value is the 'TotalCharges' variable. In assignment 1 we had postulated that total charges was the product between monthly charges and tenure. To test this, we regress TotalCharges on MonthlyCharges and tenure, and obtain a really high fit and unimodal residuals. Removing totalcharges we then see all regressors now have low VIF: multicollinearity has been reduced, the condition number goes from 2000 down to 430, which is still high in absolute terms, but a big improvement. Note that the 'categorical only' design matrix, has a condition number of only 19, but it is still desirable to keep information relative to the two independent numerical variables. Note that at this point, save for normality of errors and heteroskedasticity all OLS assumptions are satisfied, and the model still holds.

At this point we check which coefficients in the regression are not statistically different from zero (individual coefficient t-test) and find that the variables "Gender, Partner, StreamingTV, StreamingMovies, PaymentMethod" do not affect the regression, so they are removed. The regression is repeated with the new design matrix (X without the variables listed and total charges, including a constant for the intercept). From this, we finally get an idea of what variables influence the probability that a customer churns. In particular we find:

```
const : 0.35
```

```
SeniorCitizen : 0.052
Dependents : -0.022
tenure : -0.005
PhoneService : -0.131
MultipleLines : 0.014
InternetService : 0.051
OnlineSecurity : -0.047
OnlineBackup : -0.031
DeviceProtection : -0.023
TechSupport : -0.048
Contract : -0.039
PaperlessBilling : 0.051
MonthlyCharges : 0.004
```

We then create a train/test split at 75/25, and obtain a test accuracy of only 25.5%, however this is to be expected given what we are using the model for.

Given the test accuracy is so low, it is not worthwhile to interpret the significance of the regression coefficients.

3. Classification: Logistic Regression

We finally move onto the model which we find yields satisfactory test results. As with the regression case, we begin by taking the design matrix without multicollinearity using the aforementioned procedure. What we find is that the non-statistically signficiant variables are: "Gender, Partner, DeviceProtection StreamingTV, StreamingMovies, PaymentMethod". Which were those found for the linear model with the exception of 'DeviceProtection'. Again, by removing multicollinearity we make sure that the assumptions of logistic regression are not violated.

What we find in this case is similar, in the sense that residual deviations (in the context of logistic regression) are also bimodal, interestingly, when a QQ plot is done, the residuals around the positive mode seem to be normally distributed, whereas the ones near the negative mode not as much. In any case, the relationships obtained from the coefficients of the log-odds are:

```
const : 0.418
SeniorCitizen : 1.275
Dependents : 0.841
tenure : 0.965
PhoneService : 0.377
MultipleLines : 1.091
InternetService : 1.285
OnlineSecurity : 0.769
OnlineBackup : 0.879
TechSupport : 0.774
Contract : 0.491
PaperlessBilling : 1.438
MonthlyCharges : 1.028
```

We see that PhoneSerivce and Contract affect churn probability only slightly, whereas PaperlessBilling, InternetService, and SeniorCitizen affect churn the most. Therefore clients with phone service and a contract are more likely to stay loyal than senior citizens, and unexpectedly, paperless billing as well as having internet access.

We then go on to split the dataset into train and test set using the same proportions, and this time obtain a high test accuracy of around 80%. We have 1404 correct predictions out of 1758 in the test sample, and the AUC with respect to the ROC is of 0.72.

4. Classification: Other Models

At this point it remains to compare logistic regression with other more advanced classifiers, however their complexity implies a loss of granularity with regard to interpretability. The approach is the same for all chosen classifiers: use the design matrix without multicollinearity, split the dataset into train and test, and measure test performance.

What we see so far is that Logistic Regression outperforms Nearest Neighbor Method, Support Vector Machines,Gaussian Process Regression, Decision Tree, Random Forest ans similar methods available in Scikit-Learn library.

5. Conclusion and Recommendations

According to the results, we conclude that the "Logistic Regression" yields the best results in a "high-level" business approach. As keeping the current customers is more profitable then having new customers in a business sense, we should focus on increasing the proportion of phone service users and users with a contract. So, in the short-run, users who prefer paperless billing, senior users and internet service users without a phone service can be attracted to have a phone service with a contract by proposing promotions to them.