Prediction - GLM and Regularization

Jordi Tarroch Mejón

***Abstract*—Hotels have multiple ways to increase profits using data science but is not so clear how. Studying the cancellations of the bookings based on multiple features that can be extracted from each booking is a first step to approach that problem. Cancellations reduce the hotel room efficiency if the cancellation is done too late as that room will not be filled for an exact date. Prediction models such as Logistic Regression and Elastic Net can be helpful in order to predict once a booking is done if it has higher or lower odds of being canceled. Helping this way the Management of the Hotel to take the right decisions like overbooking and incorporating that into their booking systems, either online or offline.**

***Index Terms*—elastic net, glm, prediction, lambda, alpha, searchgrid, cost function, roc curve, auc, confusion matrix**

CONTENTS

I Introduction 1

II Modeling 1
- II-A Logistic Regression 1
  - II-A1 Train - Test Data Set 1
  - II-A2 Optimal Cut-Off 2
  - II-A3 Training and Testing the Model 2
- II-B Regularization 2
  - II-B1 Train - Test Data Set 3
  - II-B2 Optimal Cut - Off 3
  - II-B3 Training and Testing the Model 3

III Conclusion 3

IV References 3

I. INTRODUCTION

Being aware about the main issue related to bookings, a clear goal is set up on deciding either if a booking will be canceled or not. For that reason, a logistic regression model is a good solution as it will give an immediate answer to that question, giving a 'yes' or 'no'.

Regularization can be used to train models that generalize better on unseen data, by preventing the algorithm from overfitting the training data set.

Both techniques are used as a way to minimize cancellations and consequently increase the profits of the hotel with this information and models. But a winner is chosen between them in order to proceed with the final predictions The best model would be the one with the best accuracy

using the test sample.

Jordi Tarroch Mejón is with Data Science, CUNEF, e-mail: jordi.tarroch@cunef.edu

II. MODELING

A. Logistic Regression

Logistic regression is a method for fitting a regression curve, y = f(x), when y is a categorical variable.
The typical use of this model is predicting y given a set of predictors x. The predictors can be continuous, categorical or a mix of both.[1]

As the goal is to create a good predicting model, multicolineality is not an issue, different would be if an explanatory model was the purpose of this paper.

1) Train - Test Data Set: Data is partitioned in order to train the model and then test it.

The variables used by it are the following:

- Predictive Variables: an original set of 15 numeric variables and 16 categorical variables that after being encoded resulted in 98 total variables.
- Target Variable: IsCanceled column, column that represents the canceled bookings. These are defined by the ReservationStatus values as
  - A booking is canceled when IsCanceled equals to 1 and ReservationStatus takes either the values 'No-Show' or 'Canceled'.
  - A booking is not canceled when IsCanceled equals to ' and ReservationStatus takes either the value 'Checkout'.

A little study of the proportions of 'canceled bookings' and 'not canceled bookings' must be done to ensure that the training and testing data sets are a real representation of the original data set.

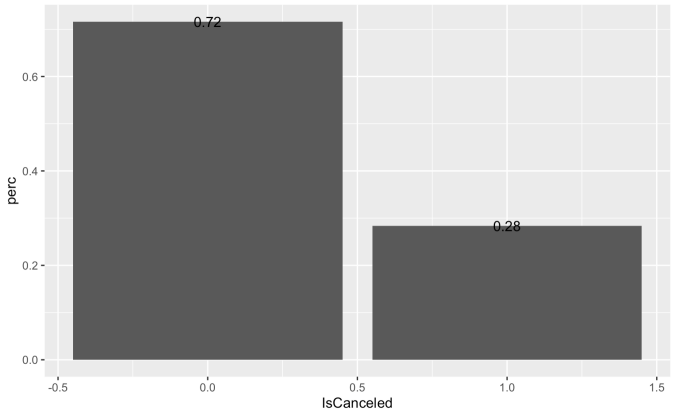

Fig. 1. Train Data Set IsCanceled Distribution

[1] https://datascienceplus.com/perform-logistic-regression-in-r/

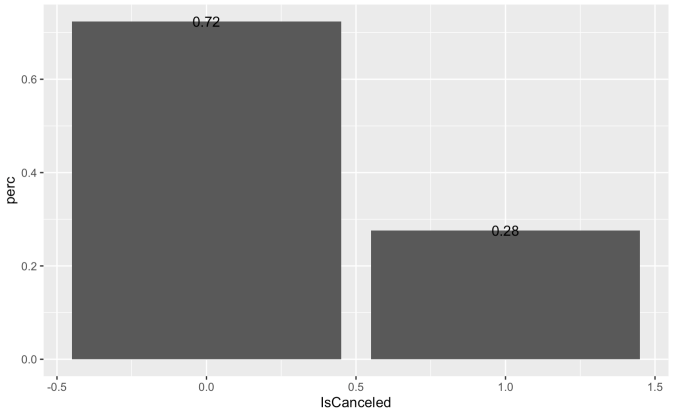

Fig. 2. Test Data Set IsCanceled Distribution

2) Optimal Cut-Off: In order to tell the model what probability should use to start distinguishing between 'Canceled' and 'Not Canceled', the cost function is defined to find the optimal cut- off that will maximize accuracy as well as find the minimum cost.

To find the minimum cost in training set the following process is followed:

- Choosing some probability cut-offs from 0.0001 to 0.6 with some increments say 0.01 and calculating the FP(False Positives) and the FN (False Negatives).
- Depending on the company costs, more weights could be given to FP or FN, which detecting them or not also depend on the cut-off probability as explained before. In this case, the same weight has been given to both.

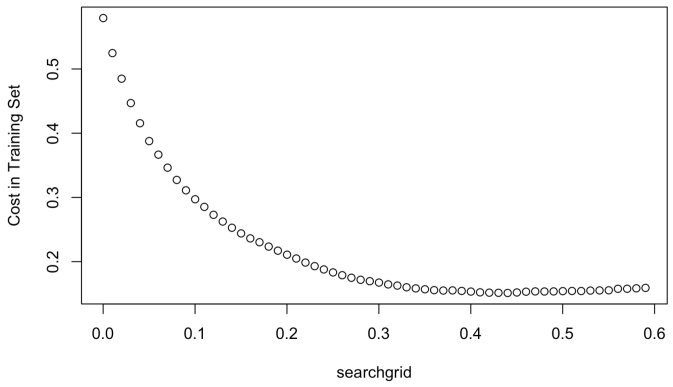

Fig. 3. Searchgrid for the Optimal Cut-Off

The Optimal Cut - Off is 0.4401000 , with a cost of 0.1512731.

3) Training and Testing the Model: The model is trained using the optimal cut-off found and then is tested with the test data set.

a) Prediction Results: To see the results a confusing matrix is used, also known as an error matrix. Is a specific table layout that allows visualization of the performance of an algorithm. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa). It is a special kind of contingency table, with two dimensions ("actual" and "predicted"), and identical sets of "classes" in both dimensions (each combination of dimension and class is a variable in the contingency table).

Based on the confusion matrix, this is how good the model performs the prediction:

- Cost of 0.1505242, similar to what was obtained in the search for the optimal cut-off.
- Accuracy of 84.87%.

TABLE I
CONFUSION MATRIX

	Predicted Not Canceled	Predicted Canceled
Truth Not Canceled	5221	517
Truth Canceled	689	1585

The Area Under the Curve is 0.9110479.

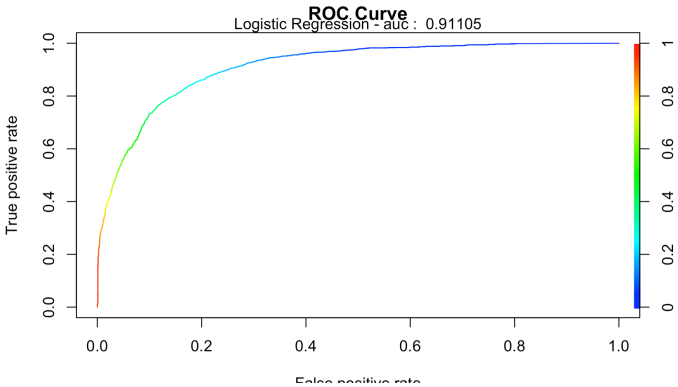

Fig. 4. ROC Curve

ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It quantifies the tradeoff made between the TPR (TruePositive Rate) and the false positive rate (FPR) at various cutoff settings (between 0 and 1).

B. Regularization

The regularized regression method used is the Elastic Net that linearly combines the L_1 and L_2 penalties of the lasso and ridge methods:

a) Lasso: is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces.[2]

b) Ridge: it is particularly useful to mitigate the problem of multicollinearity in linear regression, which commonly occurs in models with large numbers of parameters. In general, the method provides improved efficiency in parameter estimation problems in exchange for a tolerable amount of bias. [3]

c) Combination of Ridge and Lasso depending on the values of Alpha and Lambda. The best model, based on the

[2]https://en.wikipedia.org/wiki/Lasso
[3]https://en.wikipedia.org/wiki/Tikhonov

optimization of an objective function (Error, Bias, etc.), is selected via cross validation trying for different alphas and lambdas.

Hyperparemeters tuned for the Elastic Net are:

- – Alpha equal to 1 to fit with a Lasso Model.
  – Lambda that produces the highest accuracy.

Lasso Model was the method selected because of its feature selection application.

1) Train - Test Data Set: In order to train the model for the Regularization process, train and test sets are needed but with target and predictive variables separated to fit the R function with its parameters. The proportions of 'canceled bookings' and 'not canceled bookings' is the same as used for the Logistic Regression.

2) Optimal Cut - Off: Once the Lasso method is selected by adjusting Alpha equal to 1, Lambda gets selected by optimizing the accuracy of the model, finally picking the one with the greatest accuracy (lowest cost) and resulting with a Lambda equal to the 98th.

Optimal cut - off is calculated following the same process used before, but with the Lambda that gives the lowest cost.

Figure 5

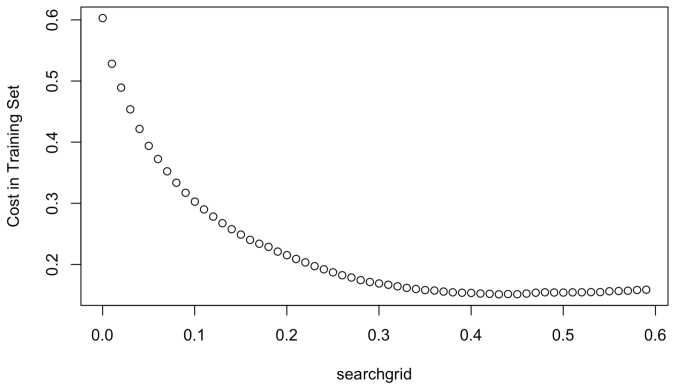

Fig. 5. Searchgrid for the Optimal Cut-Off

The Optimal Cut-Off is 0.430100, with a cost of 0.1511483.

3) Training and Testing the Model:

a) Prediction Results: Based on the confusion matrix, this is how good the model performs the prediction:

- Cost of 0.1495257.
- Accuracy of 60.02%. Table II

TABLE II
CONFUSION MATRIX

	Predicted Not Canceled	Predicted Canceled
Truth Not Canceled	5185	553
Truth Canceled	645	1629

The Area Under the Curve is 0.80999.

Figure 6

Feature Selected by Lasso: A penalty is applied over the coefficients that multiply each of the predictors. Lasso is able to give the value of zero to some of the coefficients, thus removing them from the model.

These are the features finally selected by this model:

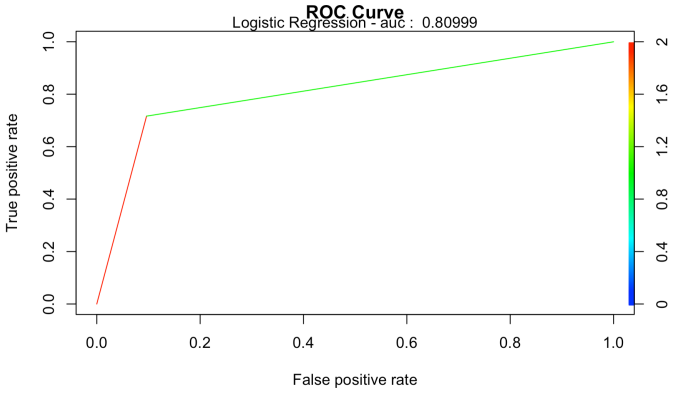

Fig. 6. ROC Curve

TABLE III
LASSO SELECTED FEATURES

Feature Selection
Deposit Type = Non Refund
Country = PRT
Market Segment = Online TA
LeadTime
RequiredCarParkingSpaces

III. CONCLUSION

The Logistic Regression has a higher Accuracy of 84.87% and better Area Under the Curve of 0.9110479 compard to the 60.02% Accuracy of the Lasso model and the 0.80999 AUC of the Lasso model. However, Lasso predicts better the Canceled bookings compared to the Logistic Regression, while the Logistic Regression predicts better the bookings Not Canceled. Depending on the costs related to the hotel, a management decision can be taken based on these results.

IV. REFERENCES

- https://datascienceplus.com/perform-logistic-regression-in-r/
- https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b
- https://www.r-bloggers.com/machine-learning-explained-regularization/