MSCA31010: Linear & Non-Linear Models

Winter 2023 Assignment 3

Since it costs more corporate resources to acquire a new customer than to retain an existing customer, companies need to preemptively identify customers who are likely to leave them. The marketing vocabulary is Churn. For a business that is based on customer subscription, churn is a serious issue. A telecommunication company provides you with the data **Telco-Customer-Churn.xlsx** and solicit your expertise to retain their customers. This data contains the following information about 7043 customers.

- Response Variable: **Churn** (the event category is *Yes*)
- Categorical Predictors: **Contract**, **Dependents**, **Gender**, **InternetService**, **MultipleLines**, **PaperlessBilling**, **Partner**, **PhoneService**, and **SeniorCitizen** (reorder the categories of each categorical predictor in ascending order of number of observations)
- Interval Predictors: **MonthlyCharges**, **Tenure**, and **TotalCharges**

You will train a binary logistic regression model with the following specifications.

- Distribution: Bernoulli
- Link Function: Logit function
- Selection Method: Backward with a Removal Criterion of 0.01 for the Deviance Significance.
- Drop all missing values (i.e., NaN) of all the predictors and the target variable.

Question 1 (20 points)

Before you train the model, you will first understand how the predictors individually affect the churn.

a) (10 points) For each categorical predictor,

- Generate a vertical bar chart that shows the odds of Churn for each category.
- Display the categories in the order of descending odds of Churn.
- Add a reference line to indicate the overall odds of Churn.
- Comment on whether it may affect the target variable.

b) (10 points). For each interval predictor,

- Generate a horizontal boxplot grouped by the target categories.
- Add a reference line to indicate the overall mean of the interval predictor.
- Comment on whether it may affect the target variable.

Question 2 (30 points)

Next, you will train your model using Backward Selection.

a) (10 points). Please provide a summary report of the Backward Selection. The report should include (1) the step number, (2) the predictor removed, (3) the number of non-aliased parameters in the current model, (4) the log-likelihood value of the current model, (5) the Deviance Chi-squares statistic between the current and the previous models, (6) the corresponding Deviance Degree of Freedom, and (7) the corresponding Chi-square significance.

b) (10 points). Please show a table of the complete set of parameters of your final model (including the aliased parameters). Besides the parameter estimates, please also include the standard errors, and the 95% asymptotic confidence intervals. Conventionally, aliased parameters have missing or zero standard errors and confidence intervals.

c) (10 points). What is the predicted probability of Churn for a customer with the following profile? Contract One year is *Month-to-month*, Dependents is *No*, Gender is *Male*, InternetService is *Fiber optic*, MultipleLines is *No phone service*, PaperlessBilling is *Yes*, Partner is *No*, PhoneService is *No*, SeniorCitizen is *Yes*, MonthlyCharges is *70*, Tenure is *29*, and TotalCharges is *1400*.

Question 3 (30 points)

You will assess the goodness-of-fit of your final model in Question 2.

a) (10 points). What is the McFadden’s R-squared, the Cox-Snell’s R-squared, the Nagelkerke’s R-squared, and the Tjur’s Coefficient of Discrimination?

b) (10 points). What is the Area Under Curve value?

c) (10 points). What is the Root Average Squared Error value?

Question 4 (20 points)

Finally, you will recommend a probability threshold for classification.

a) (10 points). Please generate the Kolmogorov-Smirnov Chart. What is the Kolmogorov-Smirnov statistic and the corresponding probability threshold for Churn? What is the misclassification rate if we use this probability threshold?

b) (10 points). Please generate the properly labeled Precision-Recall chart with a No-Skill line. According to the F1 Score, what is the probability threshold for Churn? What is the misclassification rate if we use this probability threshold?