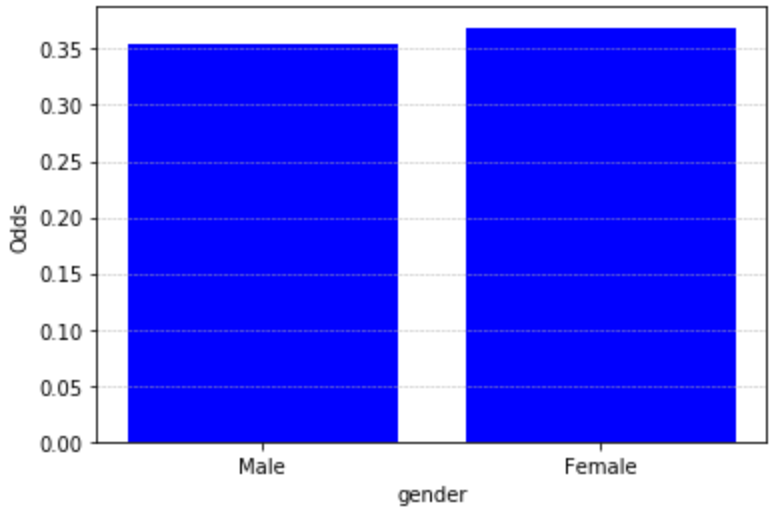
MSCA31010: Linear & Non-Linear Models

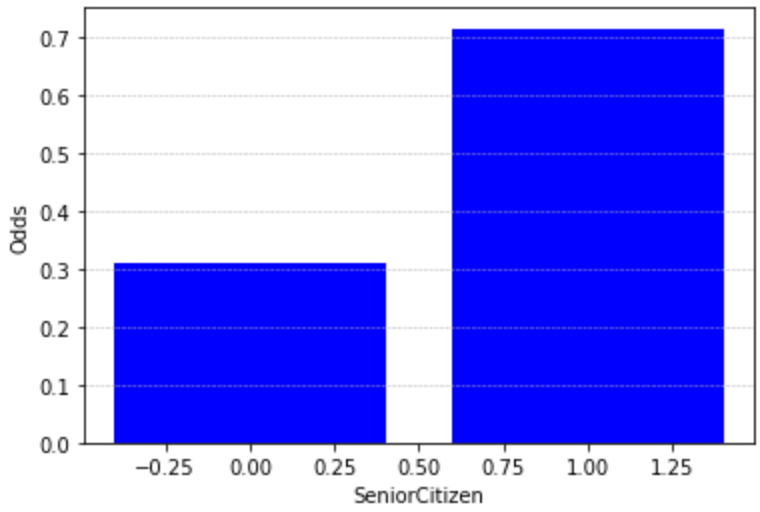
Winter 2022 Assignment 3

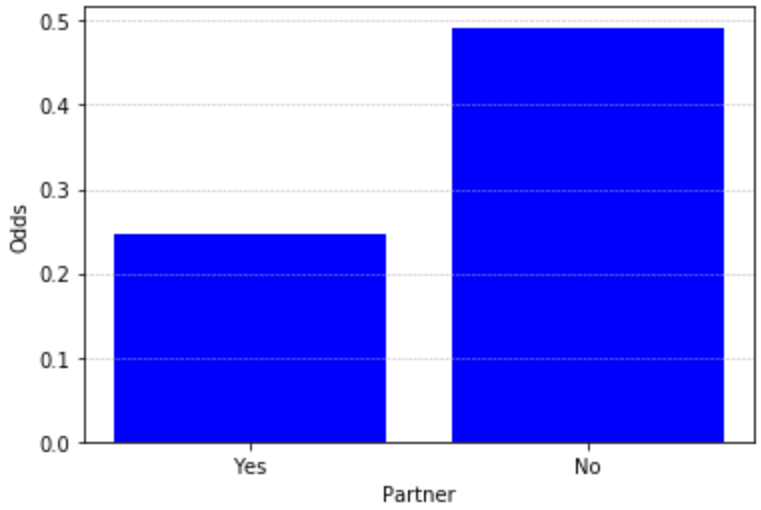
Question 1 (30 points)

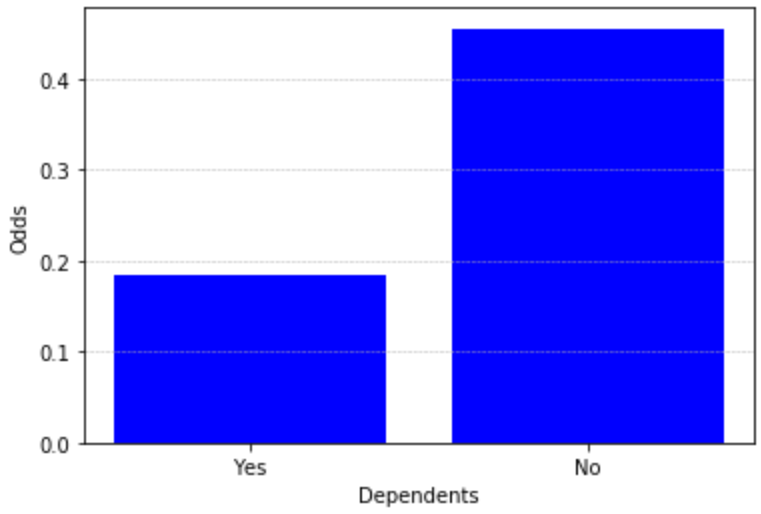
Before you train the model, you want to explore the predictors.

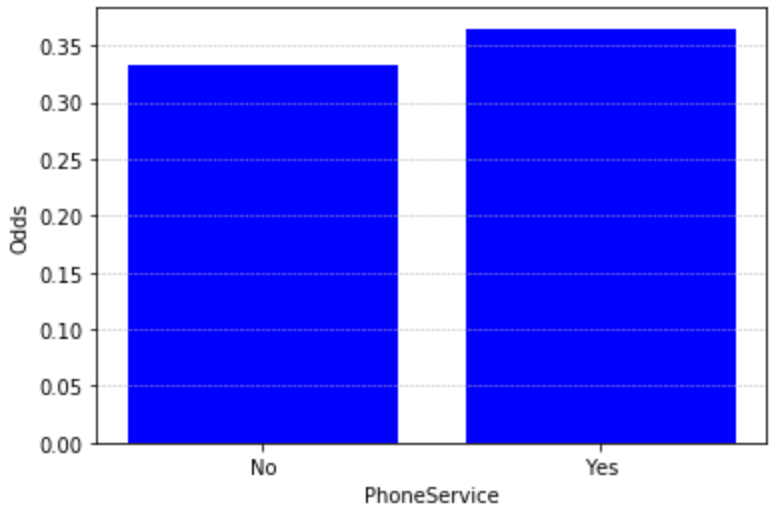
a) (20 points) For each categorical predictor, generate a bar chart that shows the odds of Churn for each category. Please order the categories in ascending odds of Churn. Also, please comment on each categorical predictor on whether it may affect the target variable.

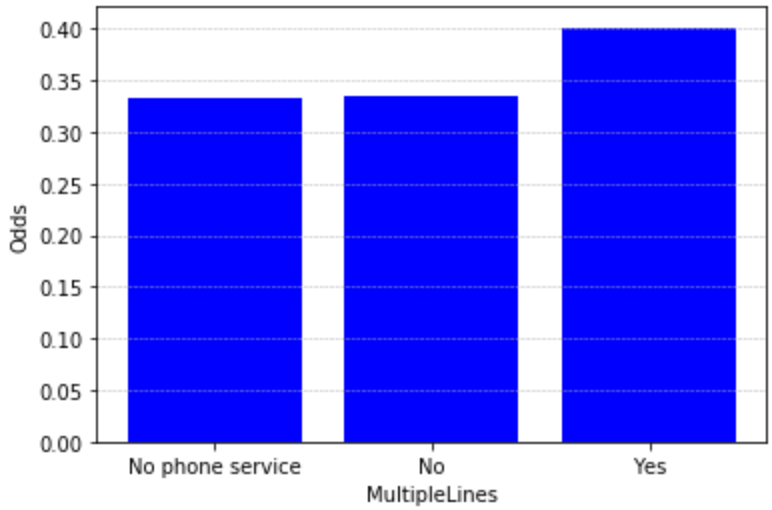


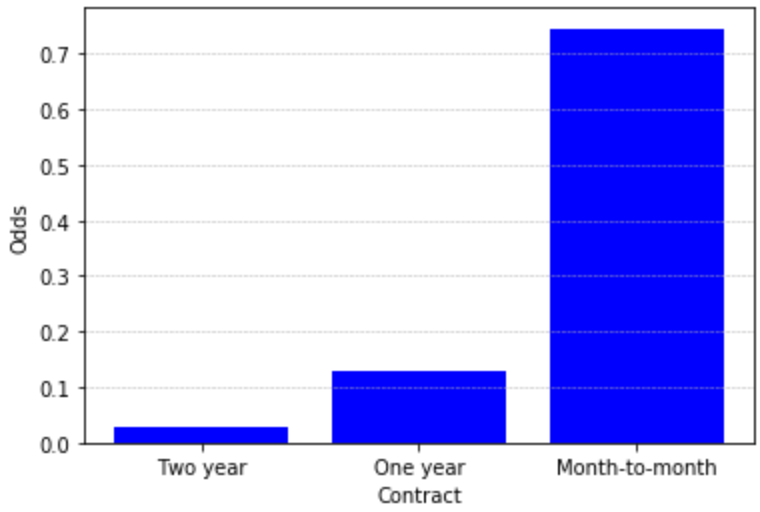


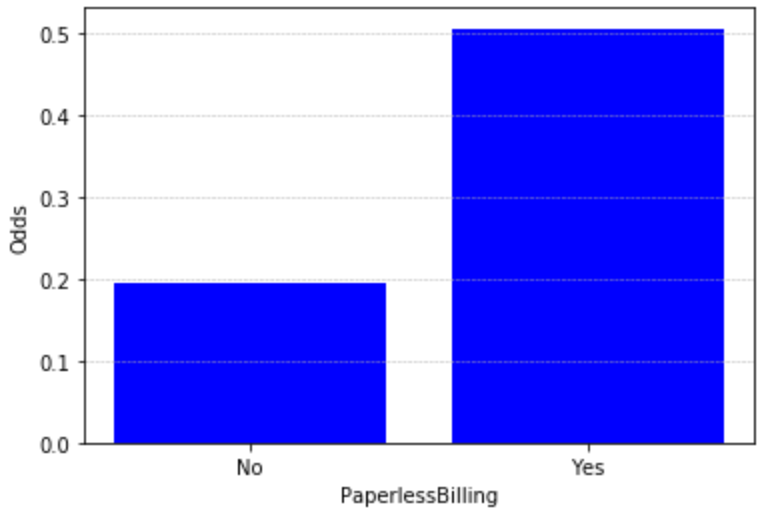

0.5
0.4
0.3
0.2
0.1
0.0
Odds
Yes
No
Partner


0.4
0.3
0.2
0.1
0.0
Odds
Yes
No
Dependents


0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Odds
No
Yes
PhoneService

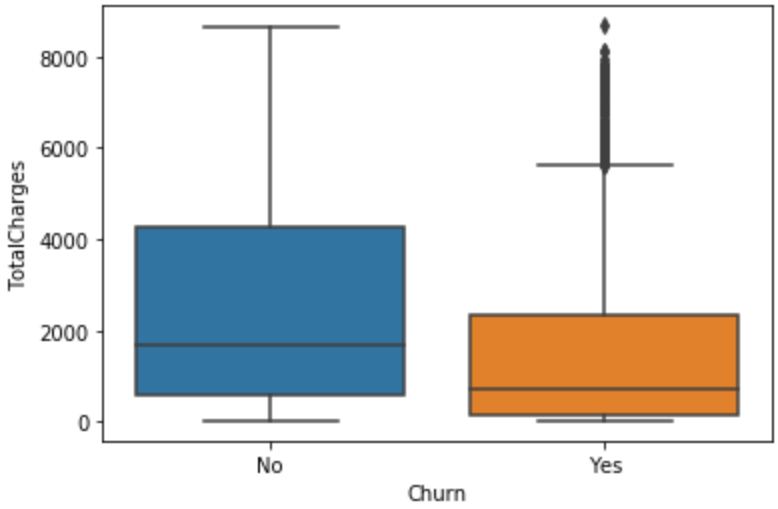

0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Odds
No phone service
No
Yes
MultipleLines

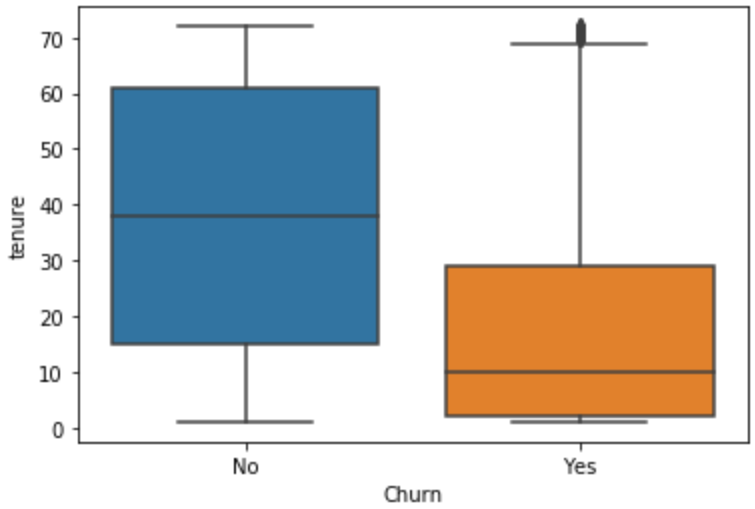

0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Odds
Two year
One year
Month-to-month
Contract

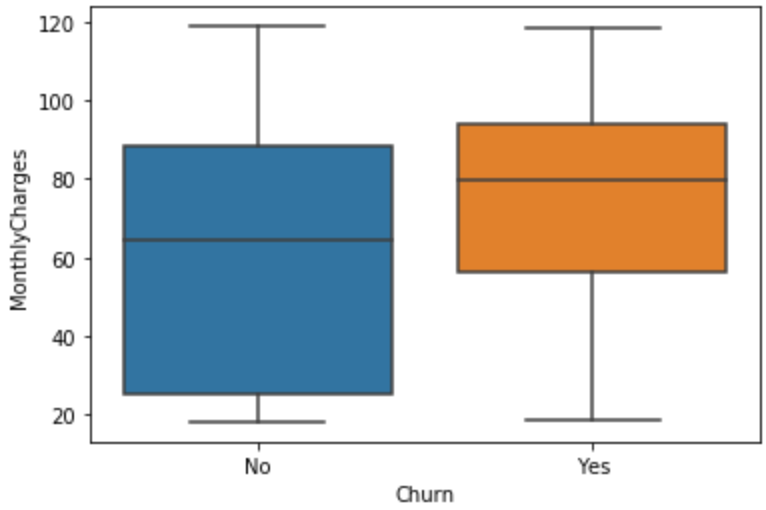

0.5
0.4
0.3
0.2
0.1
0.0
Odds
No
Yes
PaperlessBilling

From the charts, PaperlessBilling, Contract, MultipleLines, Dependents, Parter, SenitorCitizen and Gender seem to affect the output.

b) (10 points). For each interval predictor, generate a grouped boxplot that shows the distribution of the interval predictor. The grouping variable, in this case, is the target variable. Also, please comment on each interval predictor on whether it may affect the target variable.







All the three variables may affect the the target variable

Question 2 (40 points)

Enter the predictors into your model using Forward Selection. The Entry Threshold is 0.001.

a) (25 points). Please provide a summary report of the Forward Selection. The report should include (1) the step number, (2) the predictor entered, (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.

```
======= Step Summary =======
          Predictor         Type  ModelDF       ModelLLK      DevChiSq         DevDF
DevSig
0         Intercept                     1 -4.0716776e+03           nan           nan
nan
1          Contract  categorical        3 -3.3812603e+03 1.3808344e+03 2.0000000e+00 1.4308
991e-300
2    MonthlyCharges     interval        4 -3.2417829e+03 2.7895482e+02 1.0000000e+00  1.268
7061e-62
3            tenure     interval        5 -3.0717424e+03 3.4008105e+02 1.0000000e+00  6.126
5914e-76
4     MultipleLines  categorical        7 -3.0343755e+03 7.4733880e+01 2.0000000e+00  5.912
1333e-17
5  PaperlessBilling  categorical        8 -3.0152257e+03 3.8299461e+01 1.0000000e+00  6.067
9393e-10
6     SeniorCitizen  categorical        9 -3.0039965e+03 2.2458394e+01 1.0000000e+00  2.147
4496e-06
7      TotalCharges     interval       10 -2.9970235e+03 1.3946064e+01 1.0000000e+00  1.881
3117e-04
```

b) (15 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 standard errors and confidence intervals.

	Estimate	Standard Error	Lower 95% CI	Upper 95% CI
Intercept	-2.5420558e+00	2.5308454e-01	-3.0380924e+00	-2.0460192e+00
Contract_Month-to-month	1.8420787e+00	1.7108053e-01	1.5067671e+00	2.1773904e+00
Contract_One year	9.1545262e-01	1.7530432e-01	5.7186246e-01	1.2590428e+00
Contract_Two year	0.0000000e+00	0.0000000e+00	0.0000000e+00	0.0000000e+00
MonthlyCharges	2.3901287e-02	1.9943640e-03	1.9992405e-02	2.7810168e-02
tenure	-6.1276422e-02	6.0409423e-03	-7.3116451e-02	-4.9436392e-02
MultipleLines_No	-2.2265492e-01	7.9156695e-02	-3.7779919e-01	-6.7510652e-02
MultipleLines_No phone service	7.1671870e-01	1.3639833e-01	4.4938289e-01	9.8405451e-01
MultipleLines_Yes	0.0000000e+00	0.0000000e+00	0.0000000e+00	0.0000000e+00
PaperlessBilling_No	-4.3248834e-01	7.2438073e-02	-5.7446436e-01	-2.9051233e-01
PaperlessBilling_Yes	0.0000000e+00	0.0000000e+00	0.0000000e+00	0.0000000e+00
SeniorCitizen_0	-3.8308016e-01	8.0771450e-02	-5.4138930e-01	-2.2477103e-01
SeniorCitizen_1	0.0000000e+00	0.0000000e+00	0.0000000e+00	0.0000000e+00
TotalCharges	2.4983807e-04	6.7760069e-05	1.1703077e-04	3.8264536e-04

Question 3 (30 points)

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

a) (10 points). Please calculate the McFadden's R-squared, the Cox-Snell's R-squared, the Nagelkerke's R-squared, and the Tjur's Coefficient of Discrimination.

McFadden R2 = 0.26393397801947904

Cox-Snell R2 = 1.0

Nagelkerke R2 = 1.0

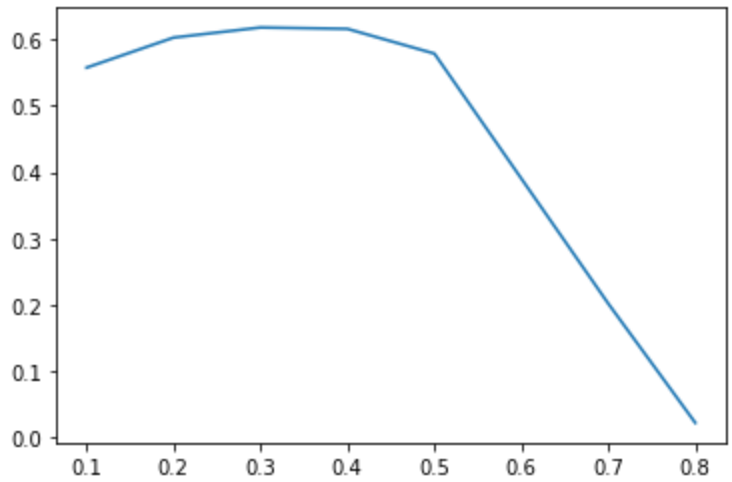
Tjur R2 = 0.28203430255131934

b) (10 points). Please calculate the Area Under Curve statistic and the Root Average Squared Error.

Area Under Curve statistic is 0.8376315

tRoot Average Squared Error is 0.3731315

c) (10 points). According to the F1 Score, please suggest the probability threshold for Churn. Using this threshold, what is the misclassification rate?



max_score=0.6178479931682322

When the threshold is 0.3, the F1 score gets the highest, so the suggested probability threshold is 0.3.

The misclassification rate is 0.254.