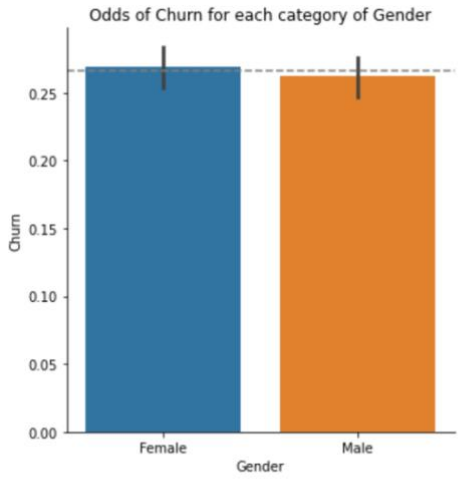
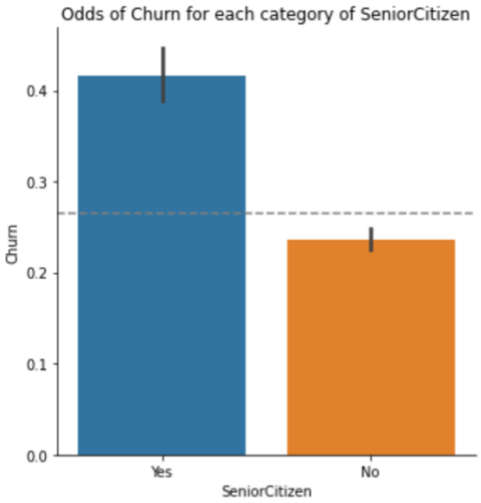
# MSCA31010: Linear & Non-Linear Models Winter Quarter 2023
## Assignment 3

### Question 1 (20 points)

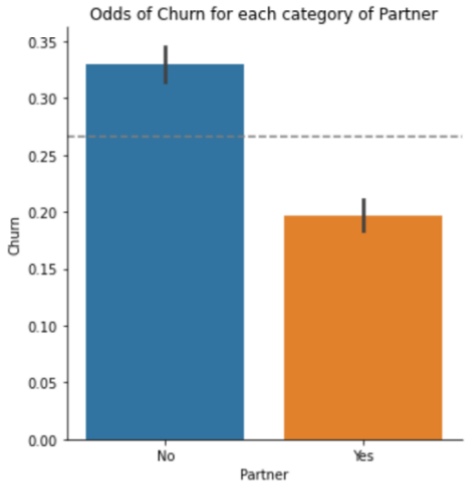
a) (10 points) For each categorical predictor,
- Generate a vertical bar chart that shows the odds of Churn for each category.
- Display the categories in the order of descending odds of Churn.
- Add a reference line to indicate the overall odds of Churn.
- Comment on whether it may affect the target variable.



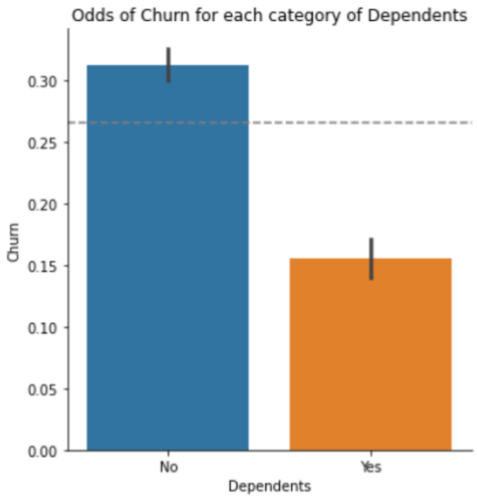

```
Odds of Churn for each category of Gender:
Female: 0.3691037735849057 (Higher than average)
Male: 0.3550973654066437 (Higher than average)
```



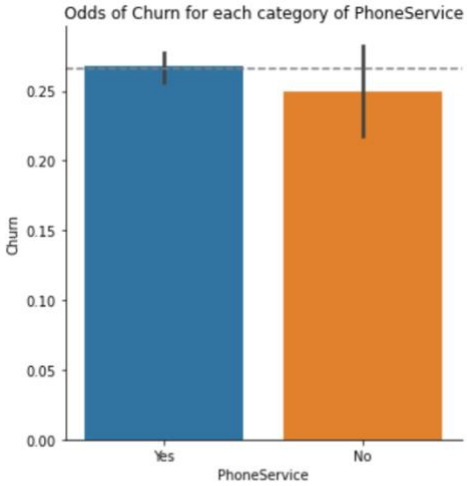

```
Odds of Churn for each category of SeniorCitizen:
Yes: 0.7147147147147146 (Higher than average)
No: 0.3097620635979542 (Higher than average)
```



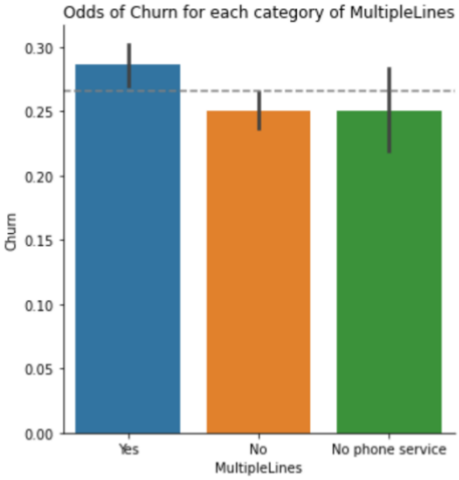

```
Odds of Churn for each category of Partner:
No: 0.49200492004920054 (Higher than average)
Yes: 0.24559471365638766 (Lower than average)
```



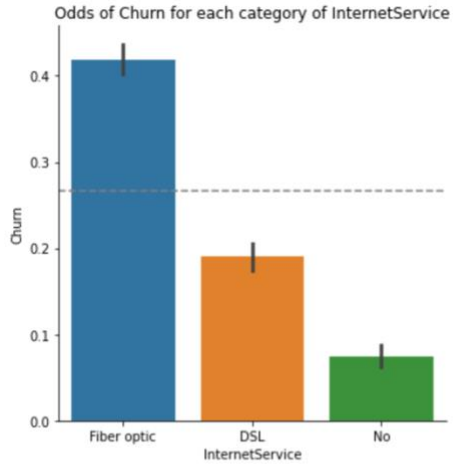

```
Odds of Churn for each category of Dependents:
No: 0.45516224188790555 (Higher than average)
Yes: 0.18386914833615342 (Lower than average)
```



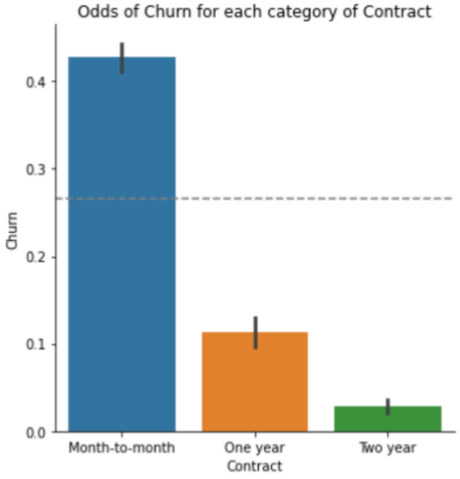

Odds of Churn for each category of PhoneService:
Yes: 0.3651407693960886 (Higher than average)
No: 0.3333333333333333 (Higher than average)



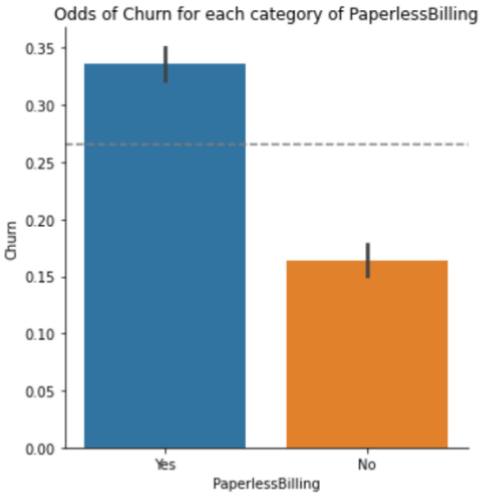

Odds of Churn for each category of MultipleLines:
Yes: 0.4015115729806329 (Higher than average)
No: 0.3347791798107256 (Higher than average)
No phone service: 0.3333333333333333 (Higher than average)




Odds of Churn for each category of InternetService:
Fiber optic: 0.7209560867148416 (Higher than average)
DSL: 0.23454266734798163 (Lower than average)
No: 0.08031272210376687 (Lower than average)



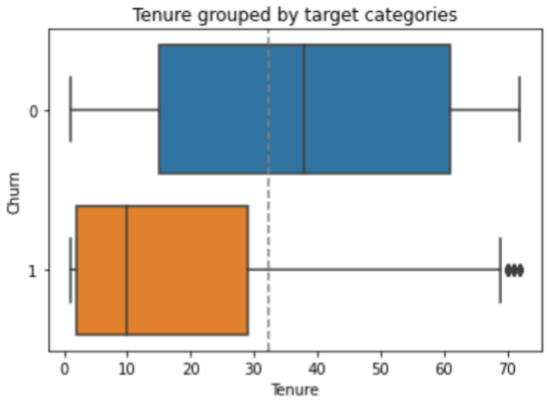

Odds of Churn for each category of Contract:
Month-to-month: 0.7454954954954955 (Higher than average)
One year: 0.12710566615620214 (Lower than average)
Two year: 0.029321930360415395 (Lower than average)




Odds of Churn for each category of PaperlessBilling:
Yes: 0.5057803468208092 (Higher than average)
No: 0.19582463465553238 (Lower than average)

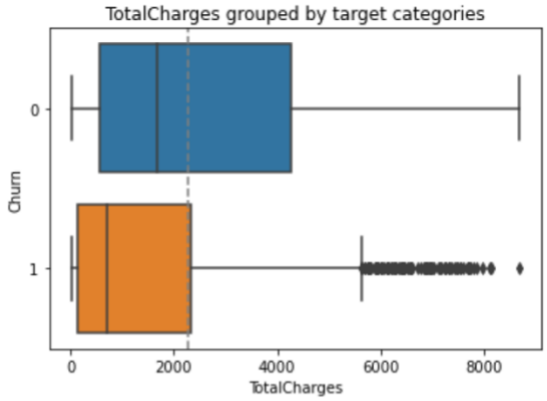
b) (10 points). For each interval predictor,

- 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.




```
Tenure mean: 32.421786120591584
May affect the target variable as it has lower values for one target category and higher
values for the other
```


```
MonthlyCharges mean: 64.7982081911263
May affect the target variable as it has higher values for one target category and lower
values for the other
```




```
TotalCharges mean: 2283.3004408418633
May affect the target variable as it has lower values for one target category and higher
values for the other
```

## Question 2 (30 points)

a) 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.

| Step | Predictor | Non-Aliased Parameters | Log-Likelihood | Deviance Chi-Squares | Degrees of Freedom | Chi-Square Significance |
|---|---|---|---|---|---|---|
| 0 | _ALL_ | 15 | -2967.432712 | NaN | NaN | NaN |
| 1 | PhoneService | 15 | -2967.432712 | 0 | 0 | NaN |
| 2 | Gender | 14 | -2967.452521 | 0.039619241 | 1 | 0.842226898 |
| 3 | Partner | 13 | -2967.474361 | 0.043678499 | 1 | 0.83445291 |
| 4 | MonthlyCharges | 12 | -2967.598075 | 0.247429445 | 1 | 0.618890928 |
| 5 | Dependents | 11 | -2969.968198 | 4.74024614 | 1 | 0.029464855 |

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.

| Estimate | Standard Error | Lower 95% CI | Upper 95% CI |
|---|---|---|---|
| 0.755165833 | 0.073488084 | 0.611131834 | 0.899199831 |
| 0.320558435 | 0.08142249 | 0.160973287 | 0.480143583 |
| 0 | 0 | 0 | 0 |
| 0.749469383 | 0.128838015 | 0.496951514 | 1.001987253 |
| 0.293111832 | 0.078419988 | 0.13941148 | 0.446812184 |
| 0 | 0 | 0 | 0 |
| -1.637992723 | 0.131830088 | -1.896374947 | -1.3796105 |
| -1.006367271 | 0.093156346 | -1.188950354 | -0.823784189 |
| 0 | 0 | 0 | 0 |
| -0.767213636 | 0.104621861 | -0.972268716 | -0.562158556 |
| -1.619149282 | 0.172939708 | -1.95810488 | -1.280193684 |
| 0 | 0 | 0 | 0 |
| -0.412285346 | 0.072908898 | -0.555184161 | -0.269386531 |
| 0 | 0 | 0 | 0 |
| -0.063021835 | 0.005877226 | -0.074540987 | -0.051502683 |
| 0.00031325 | 6.29904E-05 | 0.000189792 | 0.000436709 |

c) What is the predicted probability of Churn for a customer with the following profile? Contract One year is Month-to-mo nth, 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 1 400.

```
Predicted probability of Churn: 0.6073319944090094
```

## Question 3 (30 points)

a) (10 points). What is the McFadden's R-squared, the Cox-Snell's R-squared, the Nagelkerke's Rsquared, and the Tjur's Co efficient of Discrimination?

```
McFadden's R-squared            : 0.27057873690192324
Cox-Snell's R-squared           : 0.26899990403441476
Nagelkerke's R-squared          : 0.39218554903069747
Tjur's coefficient of discrimination : 0.2907638199477376
```

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

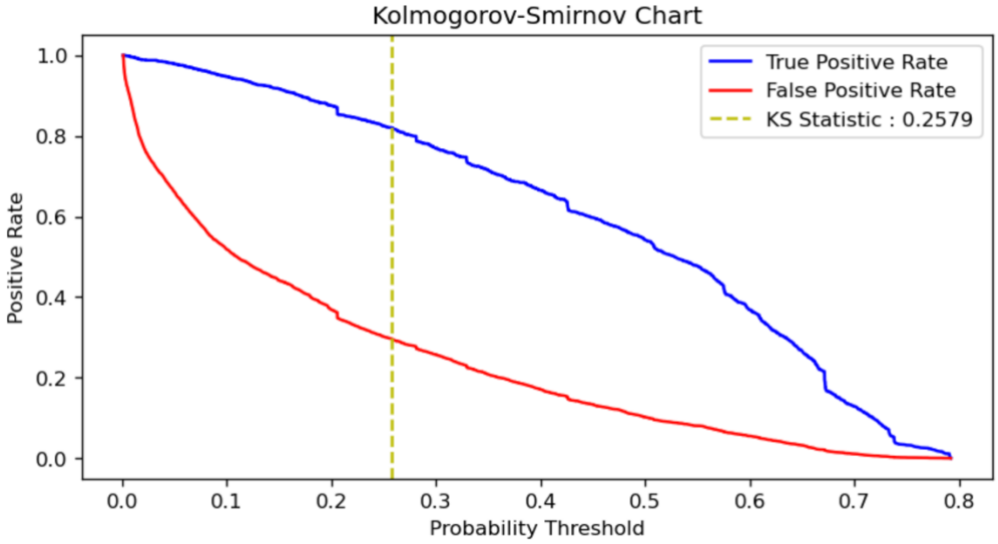
`Area Under Curve value :` **`0.8411964188949088`**

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

`Root Average Squared value :` **`0.3708047029352164`**

## Question 4 (20 points)

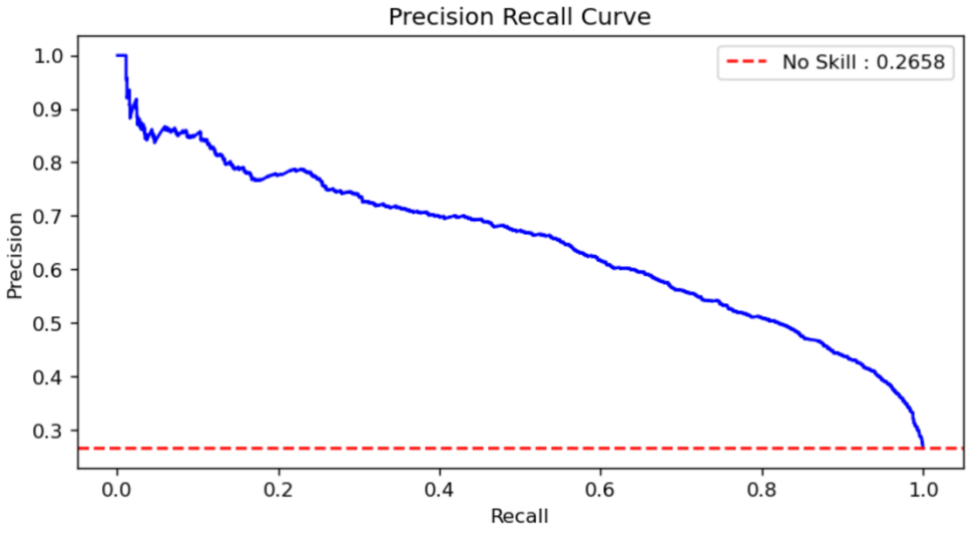
a) (10 points). Please generate the Kolmogorov-Smirnov Chart. What is the Kolmogorov-Smirnov statistic and the corresponding probability threshold for Churn? What is the misclassification rate if we use this probability threshold?

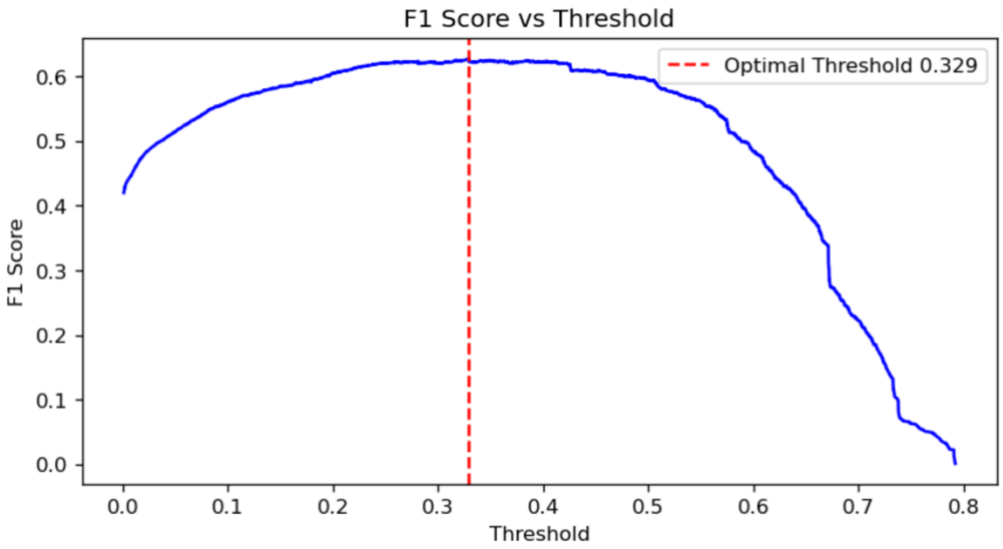



`Kolmogorov-Smirnov statistic :` **`0.5244664390313967`**
`Corresponding probability threshold for Churn :` **`0.2579355595500369`**

`Misclassification rate :` **`0.26493174061433444`**

b) (10 points). Please generate the properly labelled 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?






F-1 Score : **0.6274157303370785**
Corresponding probability threshold for Churn : **0.32902080792001165**

Misclassification rate : **0.23577929465301478**