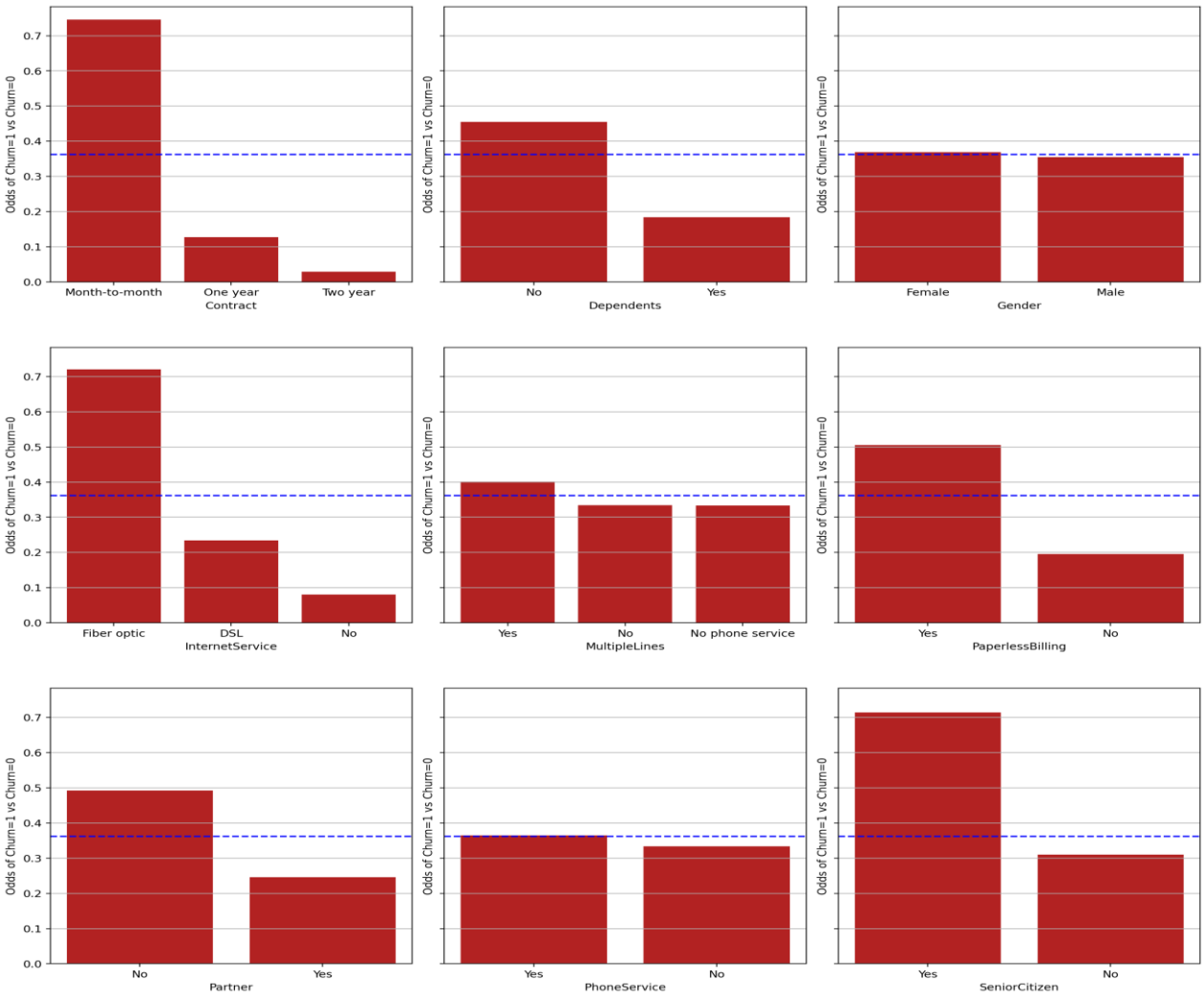
ASSIGNMENT 3

QUESTION 1

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**.



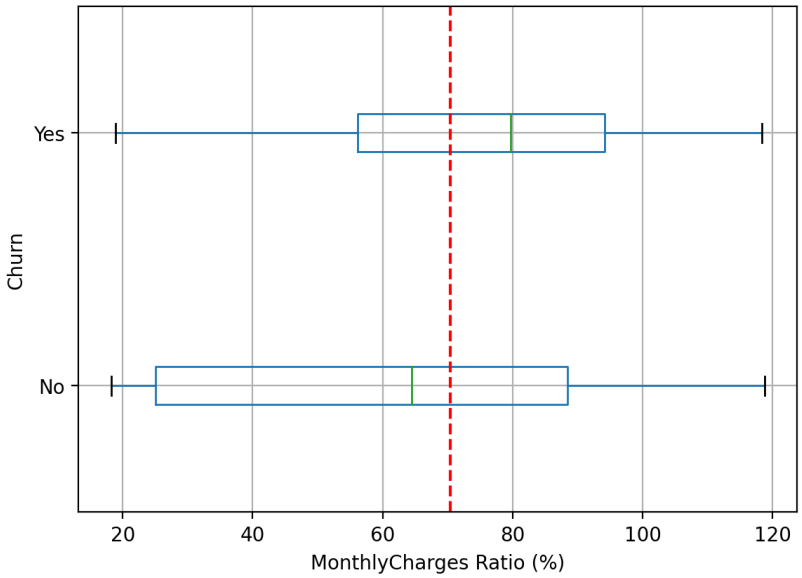
As Contract, Internet Service and SeniorCitizen have categories much higher than the average and will be assumed to have a high impact on the target variable.

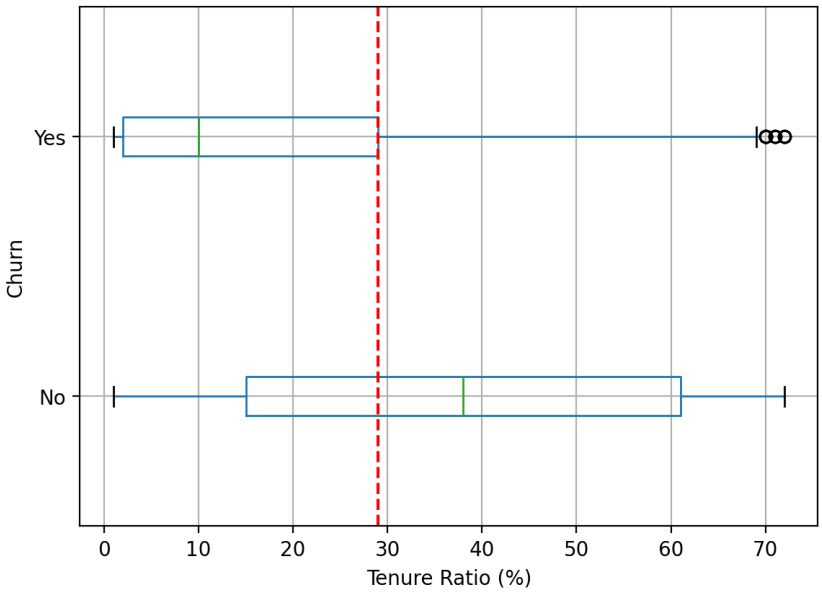
Whereas Dependents, MultipleLines and Partner have some categories somewhat above the average so it may have a smaller impact on the target.

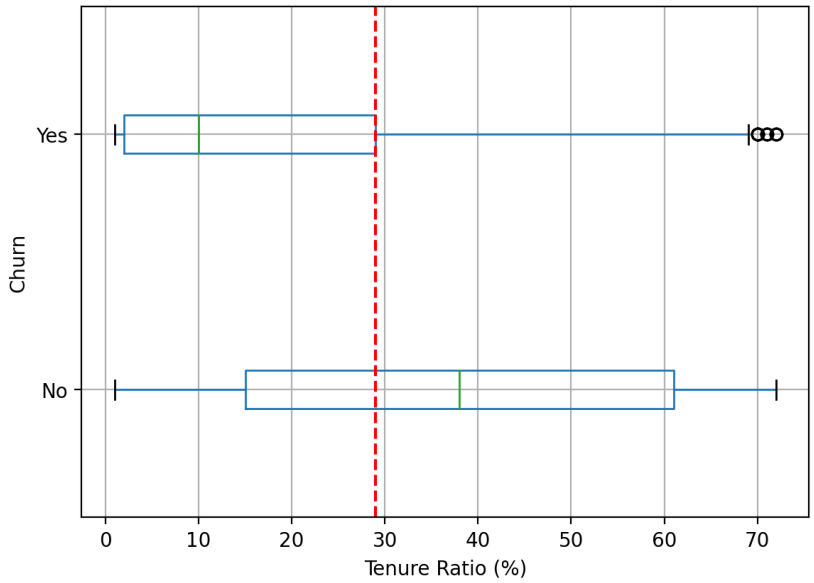
I believe Gender and PhoneService will have no impact on the target variable.

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.**







MonthlyCharges ratio has events occuring both slightly above and below average and will not have much impact on the target. Whereas, Tenure has a high chance of affecting the target because it's events (No) are higher than average even thoug it has outliers. TotalCharges on the other hand will have a medium impact on the target as it's events are slightly above average but it has many outliers present for Yes.

QUESTION 2

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.

Step	Predictor Removed	# of Non-Aliased Parameters	Log-Likelihood	Deviance Chi-Square	Deviance Degree of Freedom	Chi-square significance
0	_ALL_	15	-2967.43	NaN	NaN	NaN
1	- Gender	14	-2967.45	0.039619	1	0.842227
2	- Partner	13	-2967.47	0.043678	1	0.834453
3	- MonthlyCharges	12	-2967.6	0.247429	1	0.618891
4	- Dependents	11	-2969.97	4.740246	1	0.029465

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
Intercept	0.755166	0.073488	0.611132	0.8992
One year_Contract	-0.76721	0.104622	-0.97227	-0.56216
Two year_Contract	-1.61915	0.17294	-1.95811	-1.28019
Month-to-month_Contract	0	0	0	0
No_InternetService	-1.63799	0.13183	-1.89638	-1.37961
DSL_InternetService	-1.00637	0.093156	-1.18895	-0.82378
Fiber optic_InternetService	0	0	0	0
No phone service_MultipleLines	0.749469	0.128838	0.496952	1.001987
Yes_MultipleLines	0.293112	0.07842	0.139411	0.446812
No_MultipleLines	0	0	0	0
No_PaperlessBilling	-0.41229	0.072909	-0.55518	-0.26939
Yes_PaperlessBilling	0	0	0	0
No_PhoneService	0	0	0	0
Yes_PhoneService	0	0	0	0
Yes_SeniorCitizen	0.320558	0.081422	0.160973	0.480144
No_SeniorCitizen	0	0	0	0
Tenure	-0.06302	0.005877	-0.07454	-0.0515
TotalCharges	0.000313	0.000063	0.00019	0.000437

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.

The predicted probability of Churn for a customer with above profile = 0.6073319944090095

QUESTION 3

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 Rsquared, and the Tjur's Coefficient 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 = 0.2907638199477375

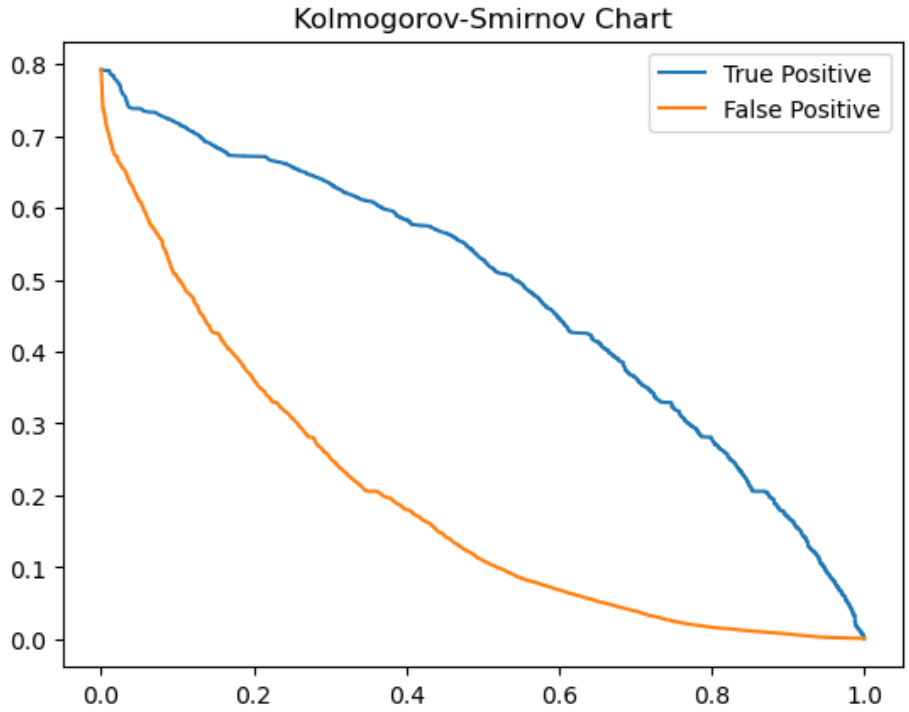
b) (10 points). What is the Area Under Curve value?
Area under the curve = 0.8411964188949088

c) (10 points). What is the Root Average Squared Error value?
Root Average Square Error = 0.37080470293521645

QUESTION 4

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?**

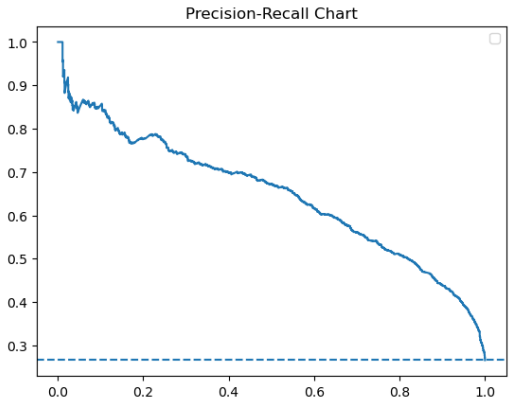


The Kolmogorov-Smirnov Statistic = 0.5244664390313967

The Corresponding probability threshold = 0.2579355595500369

The Misclassification rate for probability threshold 0.2579355595500369 = 0.26493174061433444

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?**



The F1-Score = 0.6274157303370785

The Corresponding probability threshold = 0.32902080792001154

The Misclassification rate for probability threshold 0.32902080792001154 = 0.23577929465301478