Final Capstone Presentation

By Benedict Lai

November 30, 2019

Table of Contents

	Slides
Introduction to Final Project	3
Data Analysis - Churn vs. Variables	4-13
Prediction Models	14-18
Analysis on Prediction Models	19-20
Which Prediction Model is the best?	21
Should I use Accuracy or Sensitivity?	22
Model Tuning Analysis	23-32
Unit 5 Specification Analysis	33-37
Feature Importances	38
Conclusion	39-40

Introduction to Final Project

What is the problem you are attempting to solve?

I am attempting to solve customer trends with Telco Company. For example, I want to solve if a customer is going to churn based on the tenure of the customer, the preferences for contracts and if they are a senior citizen. I mention senior citizens because they are used to paying bills through the mail and they may not be benefitting from getting a discount to keep them from staying with the company.

How is your solution valuable?

My solution is valuable because it will help determine if tenure customers are more likely to churn based on customer behavior such as the preferences for contracts, how long they have been a customer with Telco, and if they are a senior citizen. These factors may impact on customer retention. The results of my research will not only benefit Telco Company but it will also benefit other companies so it can inspire other customers to give true feedback to the company especially with the preferences of contracts.

What is your data source and how will you access it?

The data source is from https://www.kaggle.com/blastchar/telco-customer-churn. They are 7,043 customers in the dataset. I will not use the whole data set due to the null values (Total Charges group has null values) potentially impacting my project in a negative way.

What techniques from the course do you anticipate using?

I anticipate using Logistic Regression, Gradient Boosting Classifier, Random Forest Classifier, Support Vector Classifier, and Feature Importances.

What do you anticipate to be the biggest challenge you'll face?

Getting the best accuracy rate on churn, cleaning messy data, handling the class imbalance are the biggest hurdles I will face.

Data Analysis - Churn vs. Variables

Churn vs. Contract

Churn vs. Senior Citizen

Paperless Billing vs. Senior Citizen

Churn vs. Paperless Billing

Churn vs. Tenure Group

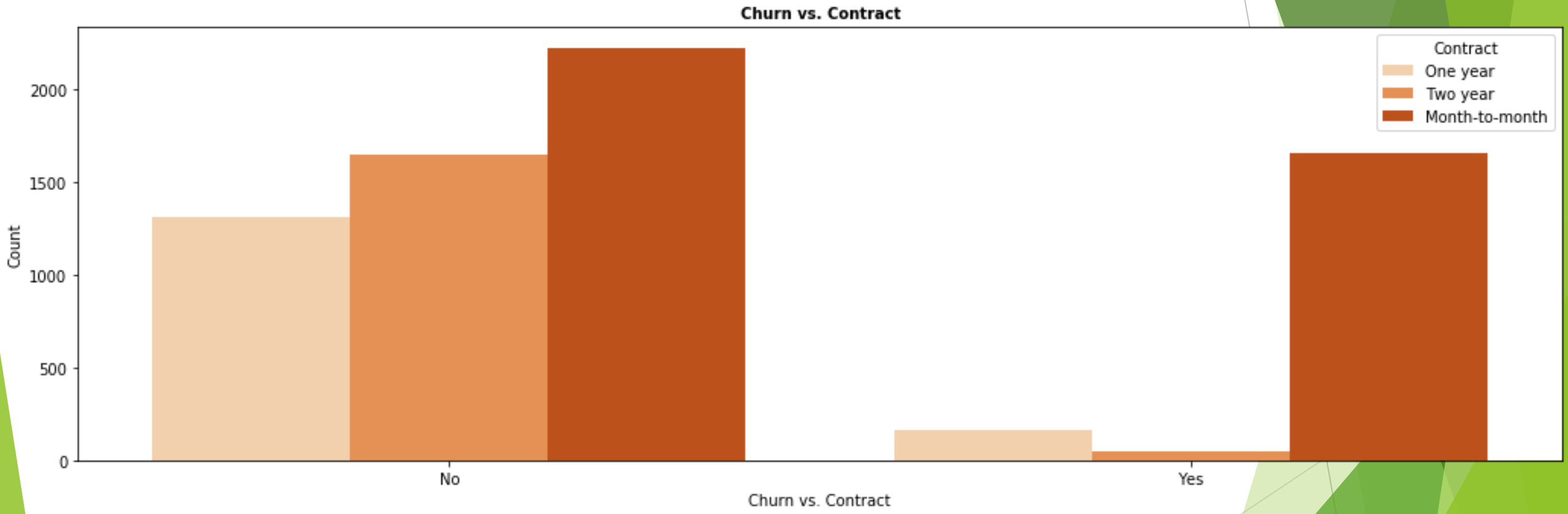
Churn vs. Monthly Charges

Churn vs. Total Charges

Churn vs. Phone Services

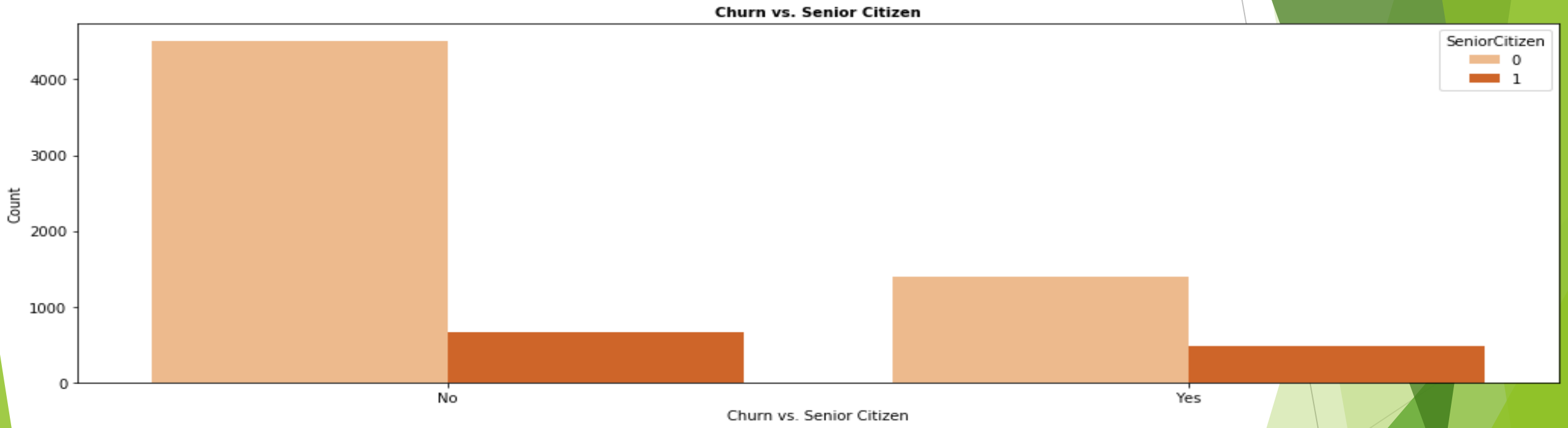
Churn vs. Internet Services

Churn vs. Contract



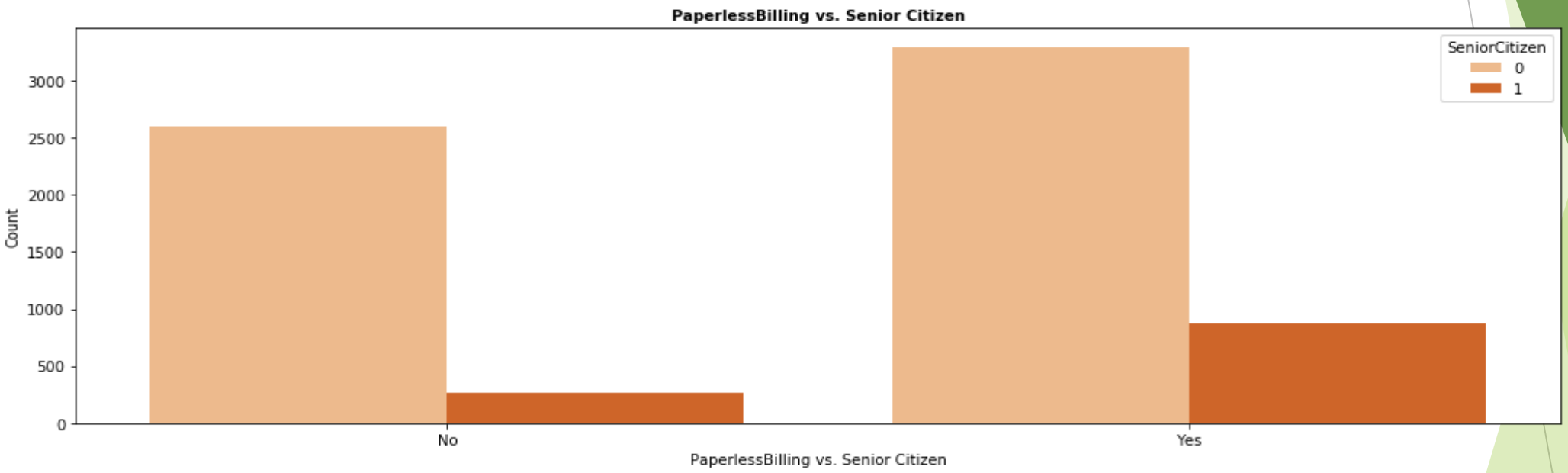
- More customers are more likely to not recommend the company with a month-to-month contract based on those who voted "Yes" to churn. A reason those types of customers only can to month-to-month contract is that they have a limited budget that prevents them from doing yearly contracts. Those types of customers may experience late fees for not paying on time, which inflates their decision to not recommend the company when the company is not at fault.

Churn vs. Senior Citizen

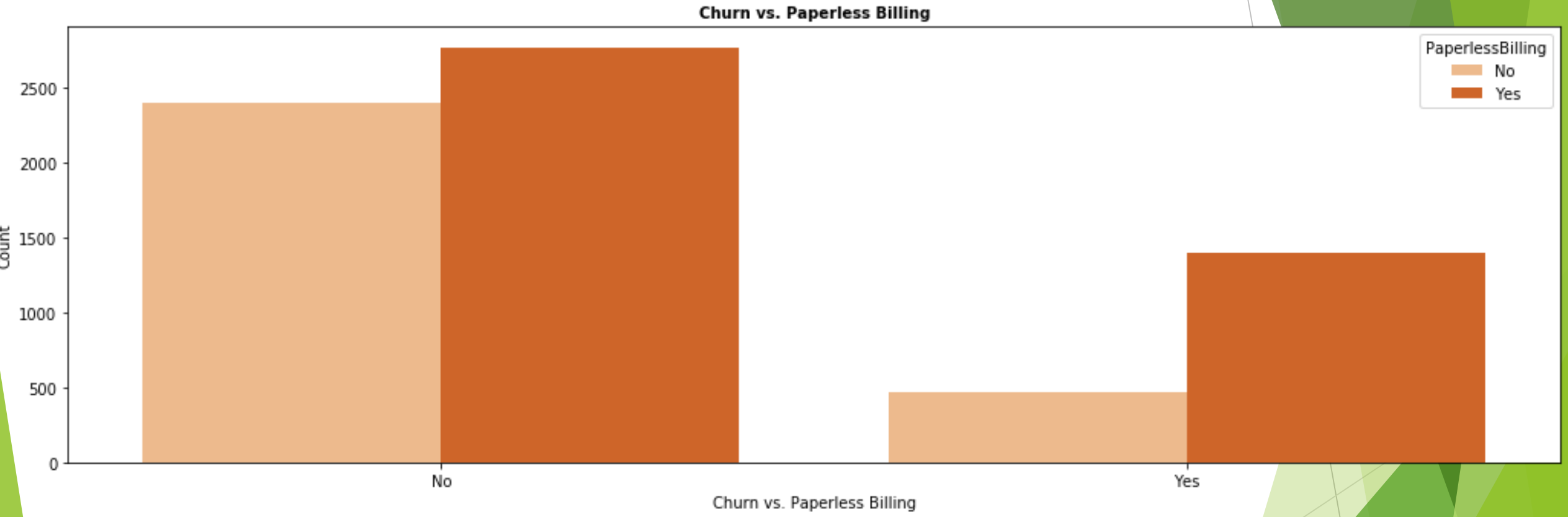


- Senior citizens are less likely to churn because there is an implication they receive a discount on services. Another implication may not be the discount, but their loyalty to the company (which causes them to have more incentives for them being with the company long). I predicted senior citizens were going to churn from the company because companies now emphasize on paperless billing. In the graph below, senior citizens did not sign up for paperless billing, which meant they are used to paying bills by mail.

Paperless Billing vs. Senior Citizen

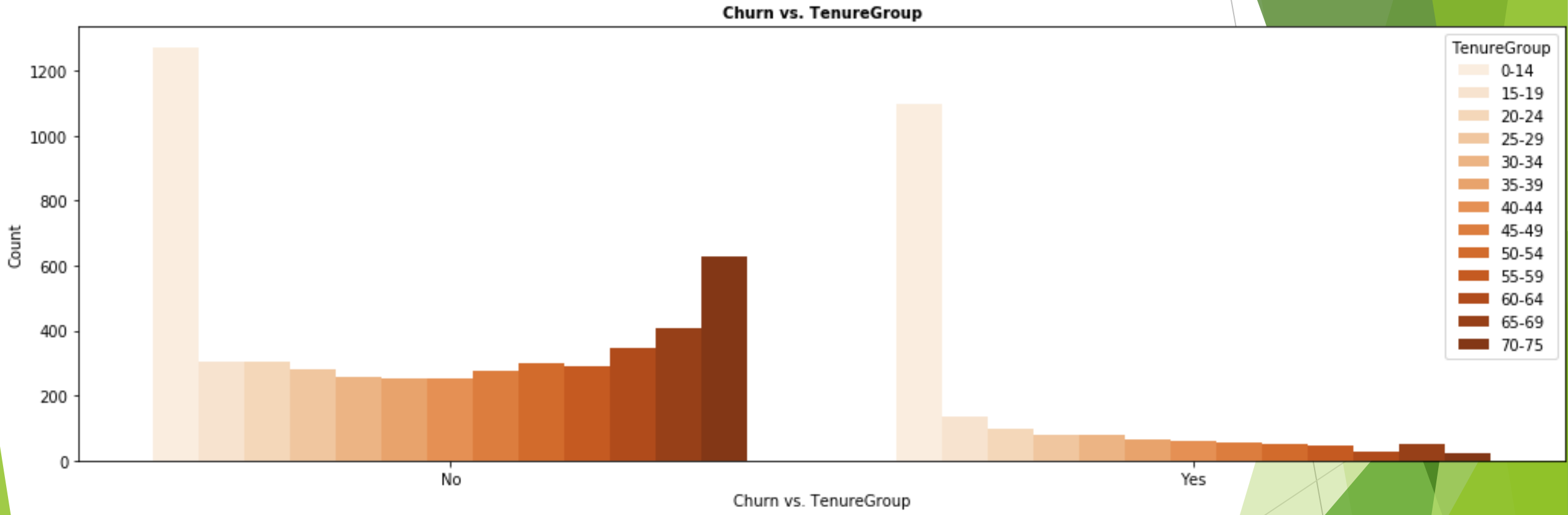


Churn vs. Paperless Billing



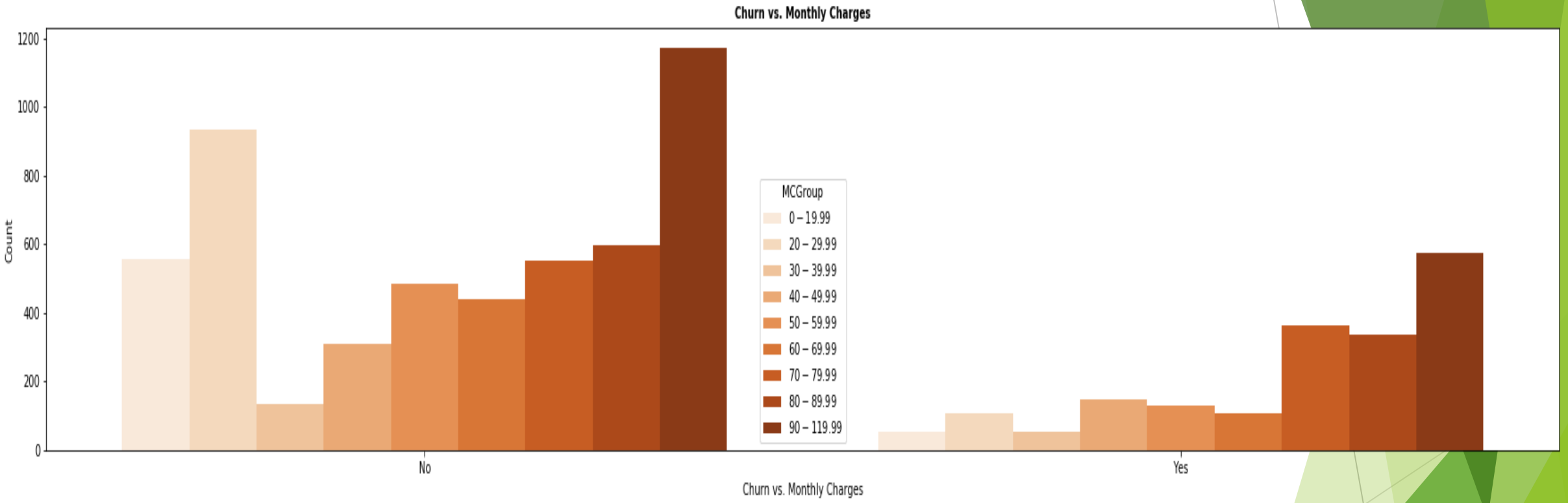
- Those who did churn actually did vote for paperless billing as their preference. What I learned from it is that not choosing paperless billing does not always play in a factor of a customer churning.

Churn vs. Tenure Group



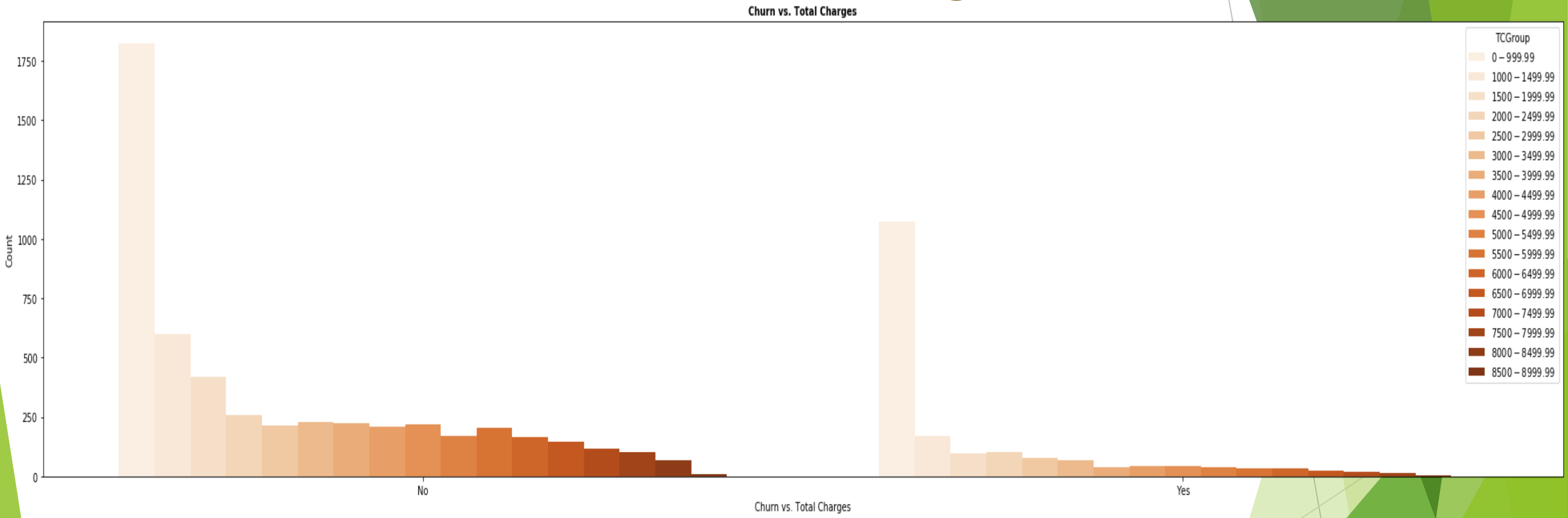
- Customers who have been with the company longer are less likely to churn because there is an implication they receive a special discount for staying with the company for a while. Those customers who have stayed with the company for less than 15 years did not recommend the company and did not show enough patience to stay longer with the company.

Churn vs. Monthly Charges



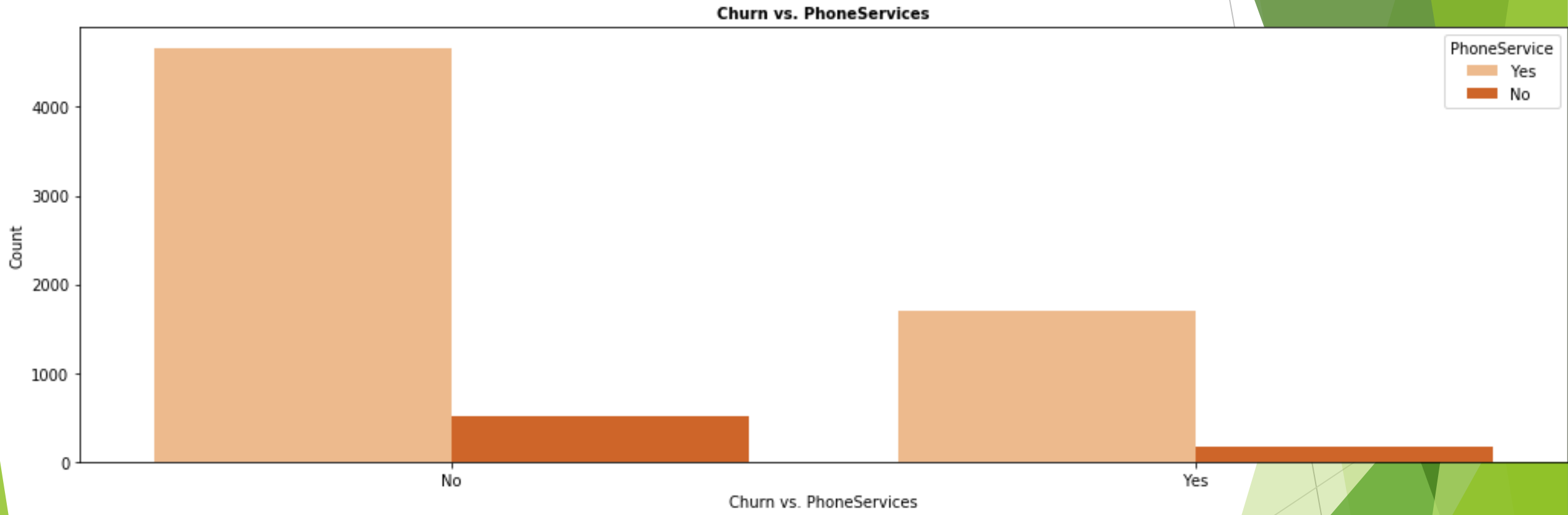
- Customers who paid $70/month are most likely to churn due to the expensive services offered. Those who paid less than $30 are satisfied with the services, which implicates they still recommended the service.

Churn vs. Total Charges



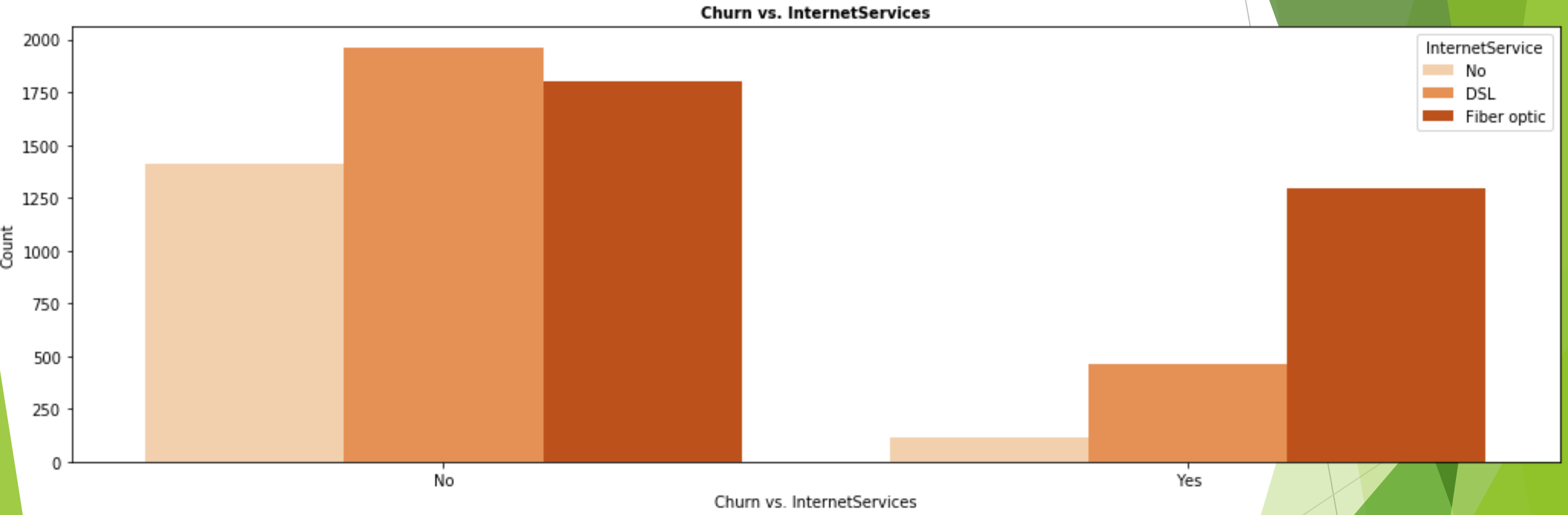
- Those churned customers are the ones that paid a total of less than $1,000 month. The more they paid in total, they less of a chance they would churn the company.

Churn vs. Phone Servies



- More customers (who churned) voted yes to having phone service. Those churned customers with phone service are unlikely satisfied with their services based on quality, and/or price.

Churn vs. Internet Services



- Those churned customers are the ones who use Fiber Optic. Because Fiber Optic is faster than DSL, it is more expensive.

Prediction Models

Random Forest Classifier

Logistic Regression

Support Vector Machine

Gradient Boosting Classifier

Random Forest Classifier

The Random Forest Classifier results are below:
Accuracy: 78.34123222748815
F1: 71.37321693055408
Precision: 73.62026466009341
Recall: 70.1157013288196
Cross Validation: [77.64920828 78.41463415 74.81707317]

Here is the confusion matrix below:
[[1347 164]
[293 306]]

Random Forest Classifier has the lowest accuracy score of 78.34% and is the lowest overall for the cross validation scores.

Support Vector Classifier

The Support Vector Classifier results are below:
Accuracy: 78.72037914691943
F1: 70.05602307297222
Precision: 75.28482294044025
Recall: 68.21400989294976
Cross Validation: [79.71985384 79.32926829 79.02439024]

Here is the confusion matrix below:
[[1398 113]
[336 263]]

Support Vector Classifier ranks 3rd in accuracy, which scored 78.72%, but it ranks 2nd in the cross validation scores.

Logistic Regression

The Logistic Regression results are below:
Accuracy: 78.86255924170617
F1: 71.32117745726183
Precision: 74.75487035272876
Recall: 69.7239718966864
Cross Validation: [79.47624848 78.90243902 77.98780488]

Here is the confusion matrix below:
[[1373 138]
[308 291]]

Logistic Regression ranks 2nd in accuracy, which scored 78.86%. However, it ranked third for the cross validation scores.

Gradient Boosting Classifier

The Gradient Boosting Classifier results are below:
Accuracy: 79.81042654028437
F1: 72.60722331119626
Precision: 76.21781351704017
Recall: 70.88960312190294
Cross Validation: [80.02436054 79.3902439 78.96341463]

Here is the confusion matrix below:
[[1383 128]
[298 301]]

Gradient Boosting Classifier has the highest accuracy score with 79.81% and is ranked 1st in cross validation scores.

Analysis on Prediction Models (Part 1)

Describe your model in detail: why you chose it, why it works, what problem it solves, how it will run in a production like environment. What would you need to do to maintain it going forward?

- I chose Random Forest because it is faster to produce results. Also, my data is not all balanced so this feature is efficient to use.
- I chose Support Vector Classifier due to its flexibility for datasets.
- I chose Logistic Regression because it is straightforward to use and easy to train.
- I chose Gradient Boosting because of the decision trees that can predict which variables has the most popularity with each of the questions. Also, it handles null values, which may be the most useful model for my dataset.
- Gradient Boosting, Random Forest, Support Vector Classifier, and Logistic Regression are the models I chose for this based on the lessons I read. I checked on all those 4 models to determine which model would fit accurately with my model.

Analysis on Prediction Models (Part 2)

Describe your model in detail: why you chose it, why it works, what problem it solves, how it will run in a production like environment. What would you need to do to maintain it going forward? (continuation)

For those models, the cross validation scores are inconsistent with the four models that I used in terms of ranking. Despite this, the models work because of the consistent scores (mainly in the 78-79 range) among the models and higher than expected scores.

The problem that it helped determine if tenure customers are more likely to churn based on customer behavior such as the preferences for contracts, how long they have been a customer with Telco, and how much they pay monthly and in total charges.

These factors may impact on customer retention. The results of my research will not only benefit Telco Company but it will also benefit other companies so it can inspire other customers to give true feedback to the company especially with the preferences of contracts.

What I would need to do to maintain it going forward is to test out one more model (which is Feature Importances) to determine which variables are the most relevant. Also, I need to do model tuning after choosing the best prediction model.

Which Prediction Model is the best?

What do the metrics you present (ie, Precision, Recall, Accuracy) mean in terms of the goals you set out for this project?

The best model is **Gradient Boosting Classifier** because they scored higher in all of these metrics than any of the prediction models. I will describe these metrics below:

76.21% of precision means it is the ratio of accurately predicted positive observations out of the overall observations (True Positive divided by True Positive + False Positive).

For Recall (sensitivity), **70.89%** is the ratio of the accurately predicted positive observations out of the overall observations in the actual class labeled (True Positive divided by True Positive + False Negative). The recall score is lower because there are more false negative observations in the dataset.

79.81% of accuracy means (True Positive + True Negative)/(True Positive + False Positive + False Negative + True Negative)). This is the highest out of the three because there are more true negative observations.

Should I use Accuracy or Sensitivity?

Should accuracy or recall be treated higher than sensitivity? Bring up these questions and an answer in your conclusions.

- Accuracy takes precedence over recall and sensitivity because it helps me choose the best model. Yes, th ere are some inconsistencies with recall score and confusion matrix not correlating well with accuracy sc ores. For example, Random Forest Classifier ranked 2nd in the recall score category and has a lower fals e negative score than Gradient Boosting, the best model. However, RFC has the lowest true positive and accuracy score.
- Despite the cons of accuracy I listed, I still choose accuracy for these reasons:
  - The accuracy scores are consistent (meaning none went way above 80 or way below 78) for all the other models.
  - Accuracy has the full formula and true story in every model. Therefore, all categories are expected to rank higher than every scoring category than the other prediction models.

Model Tuning Choice - Gradient Boosting Classifier

The Gradient Boosting Classifier results are below:
Accuracy: 79.81042654028437
F1: 72.60722331119626
Precision: 76.21781351704017
Recall: 70.88960312190294
Cross Validation: [80.02436054 79.3902439 78.96341463]

Here is the confusion matrix below:
[[1383 128]
[298 301]]

Model Tuning - Version A

The Gradient Boosting Classifier results are below:

Accuracy: 79.71563981042654 (-0.0948)

F1: 72.25785681284222 (-0.3493)

Precision: 76.24411988424617 (+0.0263)

Recall: 70.47074928542939 (-0.4189)

Cross Validation: [80.08526188 79.26829268 79.08536585]

(+0.0609, -0.1220, +0.1220)

N_estimators: 50 (-50)

Max_depth = 4 (+1)

Here is the confusion matrix below:

[[1388 123] [+5, -5]

[305 294]] [+7, -7]

Model Tuning - Version A (Explanation)

- For the accuracy, F1, precision and recall scores, they decreased by 0.0948, 0.3493, and 0.4189, respectively. However, the precision score trended up by 0.0263.
- The cross validation scores went down in the second column (down by 0.1220), but it went up in the first and third columns by 0.069 and 0.1220, respectively.
- Unfortunately, the downsides of the model is where false negatives went in the wrong direction (up by 7 to 305).
- On the bright side, the true positives (up by 5 to 1388) went in the right direction. For the next model (Version B), I will test if increasing the estimators back to 100, which I hope will help my model better.

Model Tuning - Version B

The Gradient Boosting Classifier for Version B results are below:

Accuracy: 79.81042654028437 (+0.0948)

F1: 72.92593623381178 (+0.6681)

Precision: 75.98649915720625 (-0.2576)

Recall: 71.34303919283076 (+0.8723)

Cross Validation: [79.90255786 79.3902439 78.7195122]

[-0.1827, +0.1220, -0.3659]

N_estimators: 100 (+50)

Max_depth = 4 (0)

Here is the confusion matrix below:

[[1374 137] [-14, +14]

[289 310]] [-16, +16]

Model Tuning – Version B (Explanation)

- For the accuracy, F1, and recall scores, they improved by 0.0948, 0.6681, and 0.8723, respectively compared to Version A.
- However, the precision score unusually went down by 0.2576. Additionally, the cross validation scores went down in the first and third column by 0.1827 and 0.3659, respectively. However, the second column went up by 0.1220.
- Also, the true positive went in the wrong direction (went down by 14 to 1374). Despite this disappointment, the false negative did trim down by 16 from 305 to 289, which went in the right direction.
- For the next model (Version C), I will test if increasing the estimators up to 150 will help my model better.

Model Tuning - Version C

The Gradient Boosting Classifier results for Version C are below:

Accuracy: 79.71563981042654 (-0.0948)

F1: 72.40569266270339 (-0.5202)

Precision: 76.127072787572 (+0.1406)

Recall: 70.67227642806398 (-0.6708)

Cross Validation: [79.6589525 78.84146341 78.90243902]

[-0.2436, -0.5488, +0.1829]

N_estimators: 150 (+50)

Max_depth = 4 (0)

Here is the confusion matrix below:

[[1384 127] [+10, -10]

[301 298]] [+12, -12]

Model Tuning – Version C (Explanation)

- Despite increasing the estimators up to 150, the scores went down in most of the categories.
- For the accuracy, F1, and recall scores, they decreased by 0.0948, 0.5202, and 0.6708, respectively compared to Version B even though the precision score went up by 0.1406. Also, the false negatives went in the wrong direction (going up 12 from 289 to 301). Additionally, the cross validation scores went down in the first two columns by 0.2436 and 0.5488, respectively. However, the third column went up by 0.1829.
- Despite the disappointment in the model, the true positive did improved by 10 from 1374 to 1384.
- An assumption of the model not meeting my high expectations is where I went over a certain limit where I cannot make the estimators too high. Increasing the estimators higher than 150 may cause the model to perform worse than the original model.
- Therefore, I will test if increasing the estimators down to 125 and reducing the max depth to 3 will help the next model (Version D) better.

Model Tuning - Version D

The Gradient Boosting Classifier results for Version D are below:

Accuracy: 80.0 (+0.2844)

F1: 72.82856134980183 (+0.4229)

Precision: 76.53946824631493 (+0.4124)

Recall: 71.07234758128759 (+0.40)

Cross Validation: [80.08526188 79.3902439 78.96341463]

[+0.4263, +0.5488, +0.0698]

N_estimators: 125 (-25)

Max_depth = 3 (-1)

Here is the confusion matrix below:

[[1386 125] [+2, -2]

[297 302]] [-4, +4]

Model Tuning – Version D (Explanation)

- This strategy paid off of decreasing the estimators to 125 and reducing the max depth back to 3, which caused the scores in all categories to go up vs. Version C. For example, the accuracy, F1, precision, and recall scores, improved by 0.2844, 0.4229, 0.4124, 0.40, respectively.
- Also, true positives (up 2 from 1384 to 1386) and false negatives (down 4 from 301 to 297) went in the right direction.
- Additionally, all the columns went up by 0.4263, 0.5488, and 0.0698 respectively.
- This is the only model that changed in every category in the same way for the scores.

Findings on Tuning Models

Why did you choose not to decrease the max depth or increase it further?

- When I realized I went to far by increasing the estimators dramatically when it hurt Version C model, the models indicate changing the max depth dramatically (regardless of increasing or decreasing) would also hurt my model.

What kind of inconsistences did you see for the models?

- Version D had lower true positives (1374 vs. 1386) and higher false negatives (297 vs. 289) than Version B even though Version D had a higher accuracy score. Also, Version B had the lowest number of false negatives and highest of true positives; yet, it had the lowest score in the precision category. Furthermore, the precision score did not trend the same way as the other category scores in Versions A-C compared to the original model.

Which version of the models are the best?

- Version D is the best for model tuning because it is the only model that reduced false negatives (down 4 to 297) and true positives (up 2 to 1386) at the same time. Also, it is the only model that has an accuracy score that improved vs. the original model. Additionally, Version D improved in every category vs. the original model and is the only model that matched (second column and third columns) or exceeded (first column by 0.0609) in cross-validation scores vs. the original model.

Unit 5 Specialization

Which gender of tenured customers are more likely to churn?

Which gender of senior citizens are more likely to churn?

0 = Female
1 = Male

Data - Which gender of tenured customers more likely to churn?

Gender (1 Year at Telco Company)
Female 0.523077
Male 0.509259
Name: Churned, dtype: float64
KstestResult(statistic=0.6947147446423894, pvalue=0.1863981742775230 6)

Gender (2 Years at Telco Company)
Female 0.495050
Male 0.444444
Name: Churned, dtype: float64
KstestResult(statistic=0.671639356718114 7, pvalue=0.2156414241129870 6)

Gender (3 Years at Telco Company)
Female 0.534091
Male 0.409091
Name: Churned, dtype: float64
KstestResult(statistic=0.6587635262502591, pvalue=0.2328846620343151 8)

Gender (4 Years at Telco Company)
Female 0.469697
Male 0.492537
Name: Churned, dtype: float64
KstestResult(statistic=0.6807142334114626, pvalue=0.2038868014920599 3)

Analysis - Which gender of tenured customers are more likely to churn?

I chose tenured customers because they are relevant to the data set with churning. I broke down by each gender and it seems Telco female customers are more likely to churn. This indicates more females are likely to call out every mistake on a bill. Also, those type of customers are savvy, meaning they wanted a cheaper service since the service did not live up to quality.

Also, for the customers that were with Telco for only one year, new customers may not have seniority and get special treatment on certain services like the tenured customers.

Because the p-values in all tenured customer groups are higher than 0.05 for tenure vs. churning, we cannot reject the null hypothesis. Therefore, the relationship between tenure and churning does not correlate well.

Data - Which gender of senior citizens are more likely to churn?

Gender
Male 0.422535
Female 0.411150
Name: Churned, dtype: float64
KstestResult(statistic=0.6595186617539919, pvalue=0.2318550833875852)

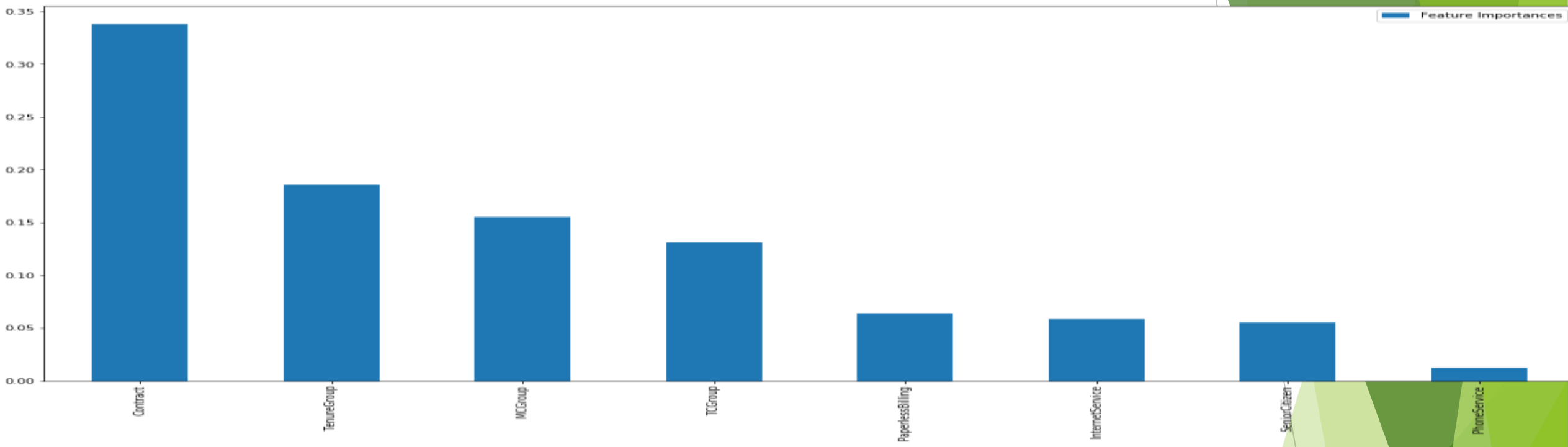
Analysis - Which gender of senior citizens are more likely to churn?

I chose senior citizen because they are relevant to the data set with churning. With the increase of paying bills online electronically, I figured this type of demographic will likely churn since they have yet to transition into the new technology era.

I broke down by each gender and it seems more female senior citizens are likely to churn Telco Company. Similar to my previous question in Unit 5, this indicates more females are likely to call out every mistake on a bill. Also, those type of customers are savvy, meaning they wanted a cheaper service since the service did not live up to quality.

Because the p-value is higher than 0.05 for senior citizens vs. churning, we cannot reject the null hypothesis. Therefore, the relationship between senior citizens and churning do not correlate well.

Feature Importances



Feature Importances	Percentage
Contract	0.338167
Tenure Group	0.185789
Monthly Charges Group	0.155077
Total Charges Group	0.130739
Paperless Billing	0.063807
Internet Service	0.058502
Senior Citizen	0.055527
Phone Service	0.012392

Contract and Tenure Group rank the highest because those groups correlate well in determining whether a customer will churn. With contracts, the majority of churned customers represent the month-to-month contract, which correlates well with customer churning. With Tenure Group, the majority of customers rank younger, which correlates well with churning Telco. I indicated those variables in the graphs earlier in slides 5 (Contract) and 9 (Tenure Group).

Senior Citizen and Phone Service ranks the lowest because they had a lower than expected number in customer churning. I indicated those variables in the graphs earlier in slides 6 (Senior Citizen) and 12 (Phone Service).

Conclusion (Part 1)

Based on your modeling and analysis, what recommendations would you be able to confidently give in order to prevent a customer from churning?

The recommendations to prevent from a customer from churning are the following:

- Meet or exceed customer expectations
- Turn weaknesses into strengths
- Offer more discounts for certain services and less tenured customers
- Emphasize on customers' complaints
- Improve communication skills with the customer(s)
- Look out for other competition and match or better their competitors' offer for that particular competition

Conclusion (Part 2)

What you set out to do?

- If the duration of a customer staying with the company, the total and monthly charges paid by the customer, the type of the contract, internet service, phone service, and the preference of paperless billing for the customer and the customer's age would be good indicators of customer churning.

What you learned along the way?

- I learned that just because customers take advantage of paperless billing and pay electronically can still mean they vote to churn. It's other factors that caused them to churn.

What conclusions you were able to confidently come to?

- The tenure of the customer, the type of contract and the total/monthly charges of the customer predicts accurately when it comes to customer churning.