Final Capstone Presentation

By Benedict Lai

November 21, 2019

Table of Contents

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 7,044 customers in the dataset. I will not be using the whole data set because of some of the null values (TotalCharges) that may affect it.

What techniques from the course do you anticipate using?
I anticipate using Logistic Regression, Gradient Boosting & Random Forest Classifier, Support Vector Machine, 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'll face.

Data Analysis - Churn vs. These Variables

Churn vs. Contract

Churn vs. Senior Citizen

Paperless Billing vs. Senior Citizen

Churn vs. Paperless Billing

Churn vs. Tenure Group

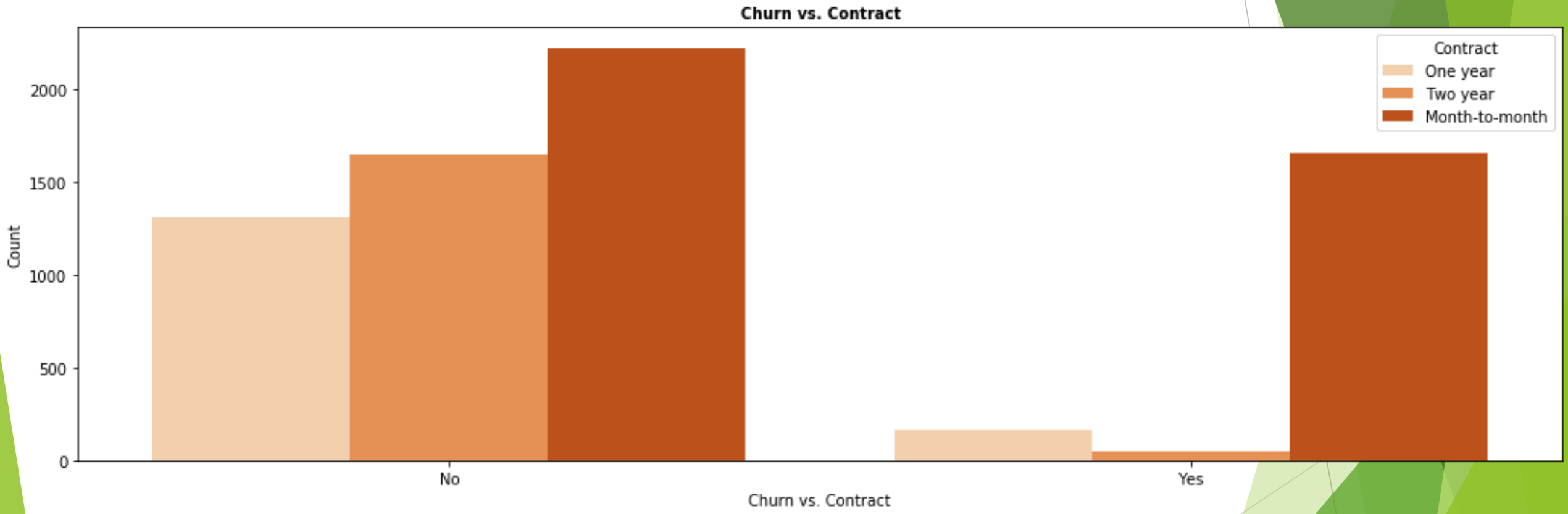
Churn vs. Monthly Charges

Churn vs. Total Charges

Churn vs. Phone Services

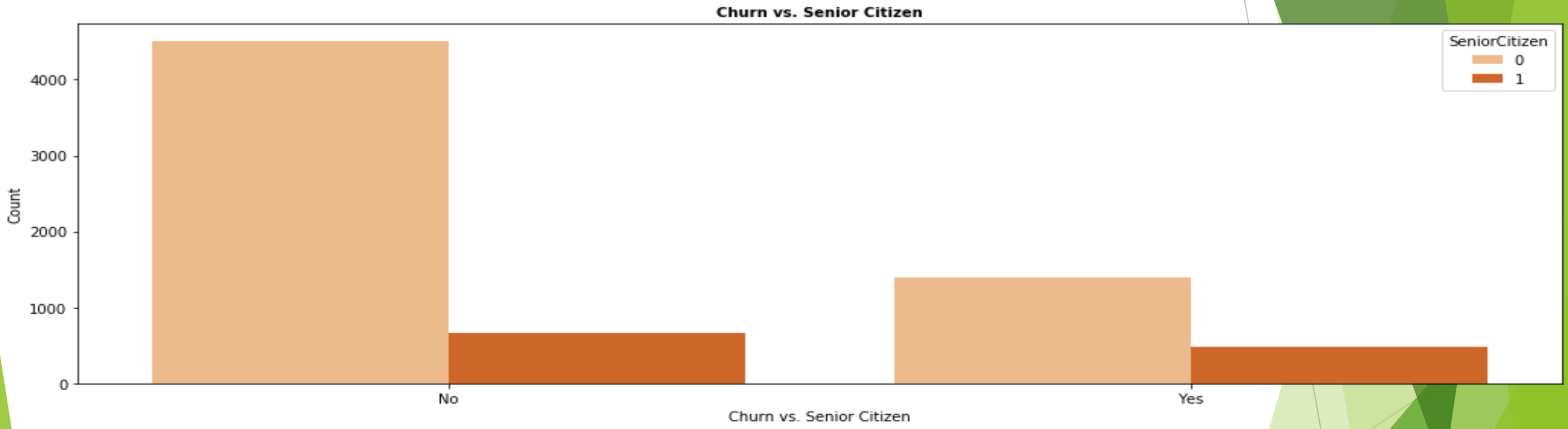
Churn vs. Internet Services

Churn vs. Contract



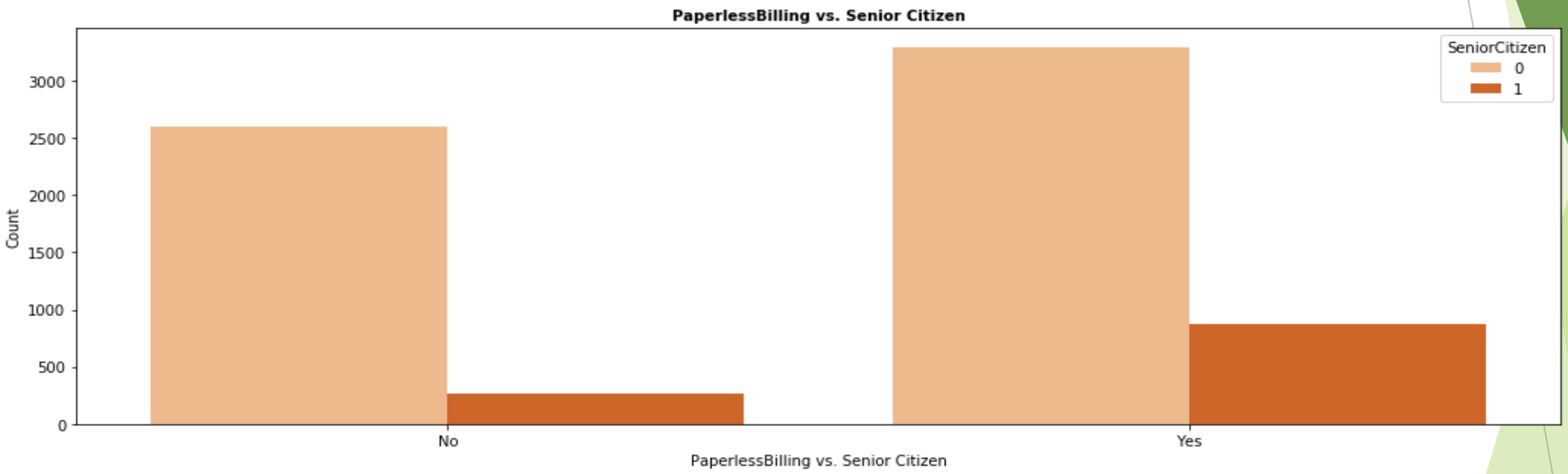
- More customers are more likely to not recommend the company with a month-to-month contract based on those who voted "Yes" to churn. A reason those types of customers only can to month-to-month contract is that they have a limited budget that prevents them from doing yearly contracts. Those types of 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

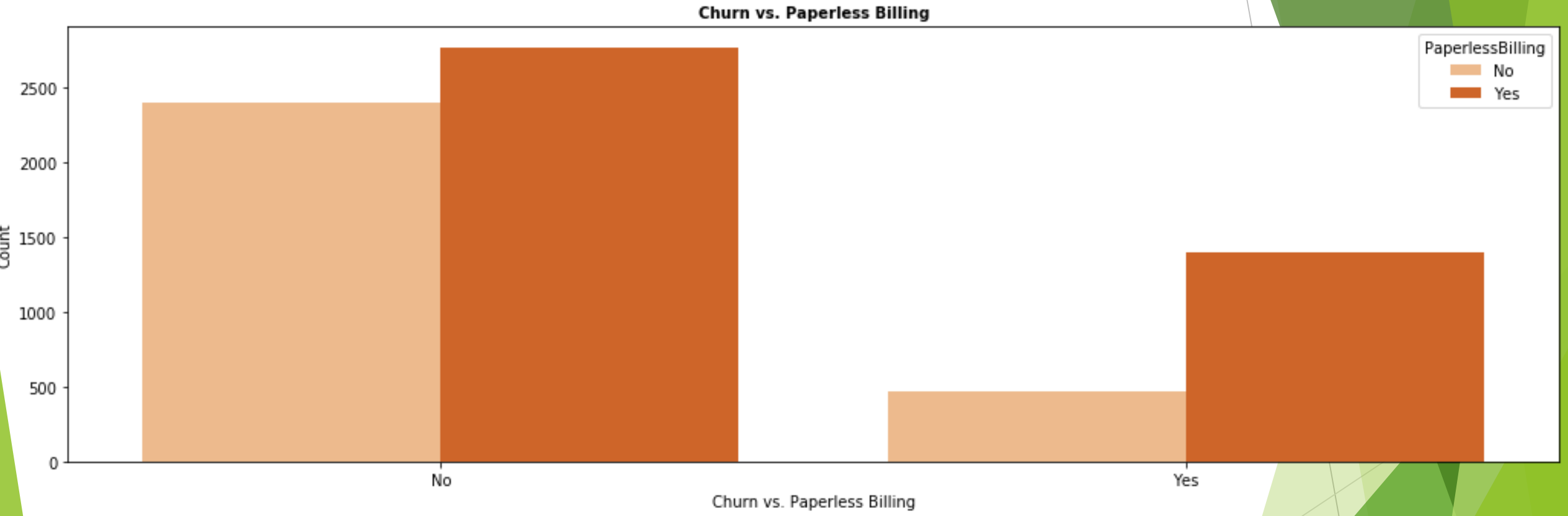


- Senior citizens are less likely to churn because there is an implication they receive a discount on services. Another implication may not be the discount, but their loyalty to the company (which causes them to have more incentives for them being with the company long). I predicted senior citizens were 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

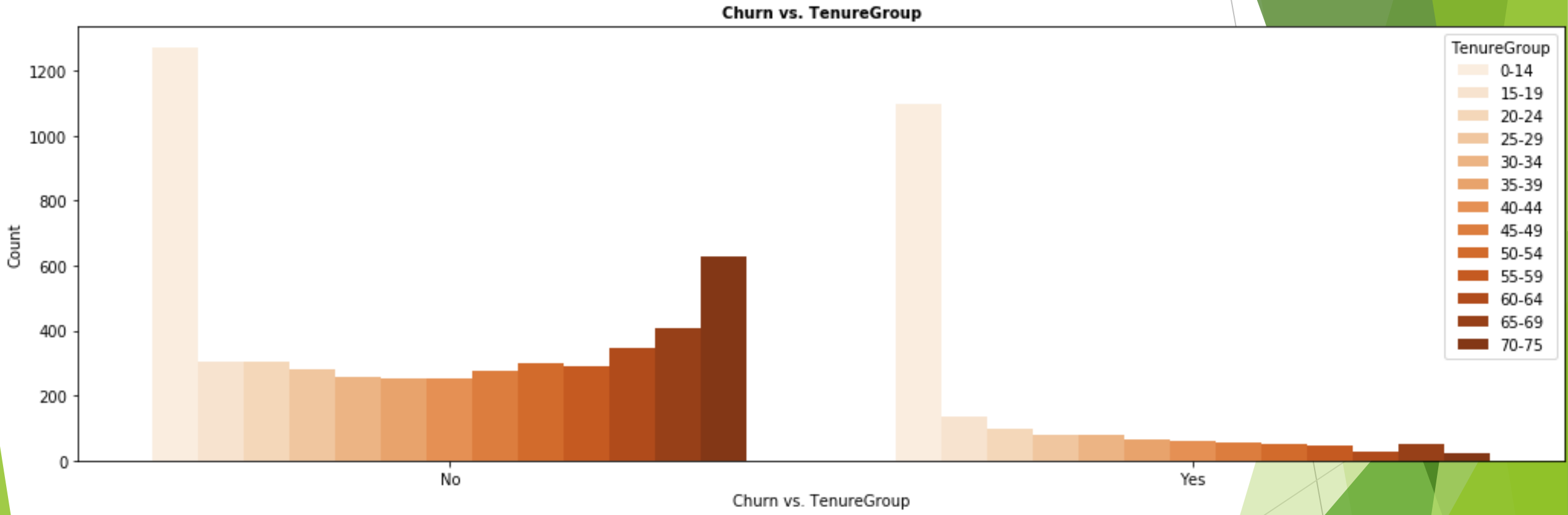


Churn vs. Paperless Billing



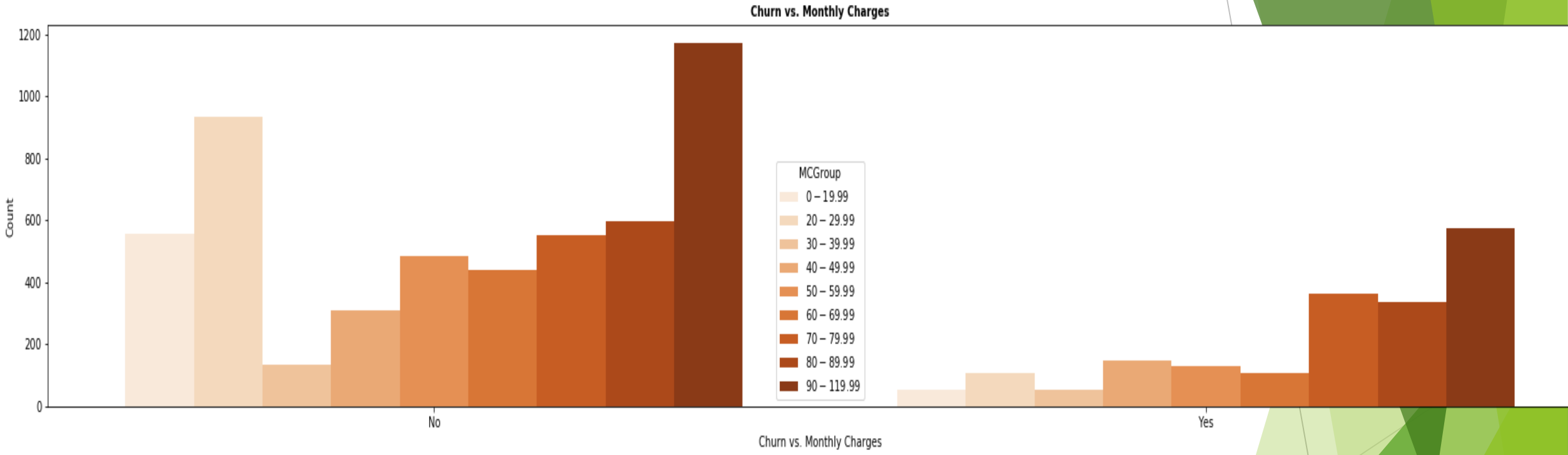
- Those who did churn actually did vote for paperless billing as their preference. What I learned from it is that not choosing paperless billing does not always play in a factor of a customer churning.

Churn vs. Tenure Group



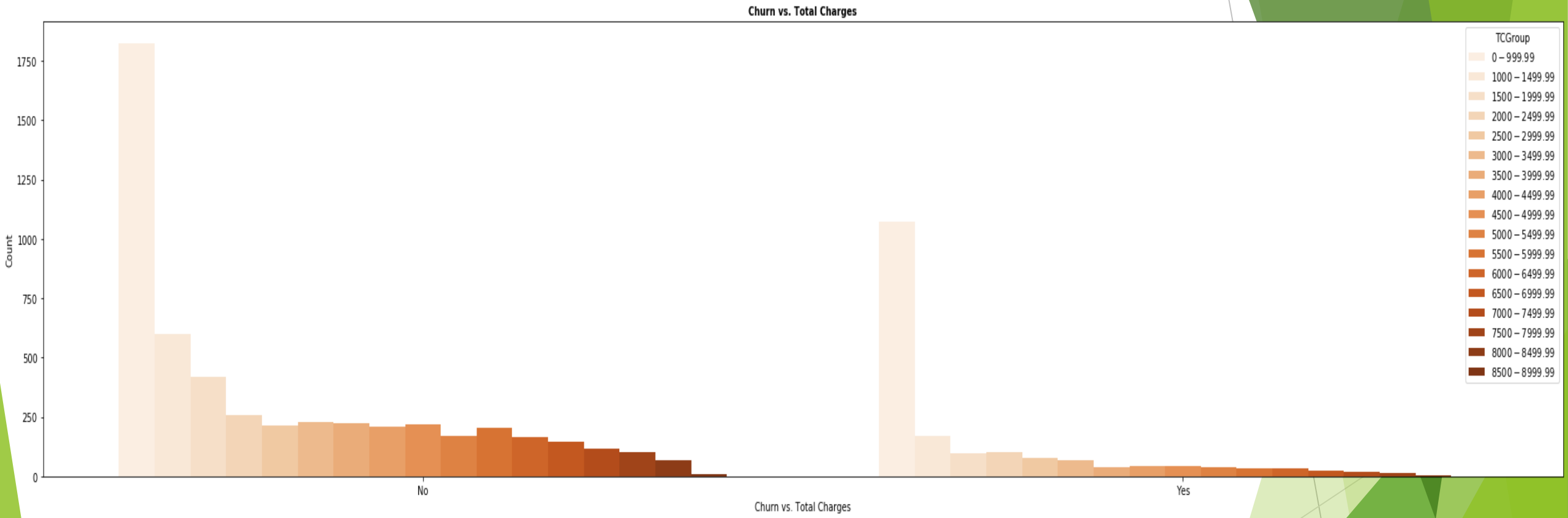
- Customers who have been with the company longer are less likely to churn because there is an implication they receive a special discount for staying with the company for a while. Those customers who have stayed with the company for less than 15 years did not recommend the company and were not patient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

Churn vs. Monthly Charges



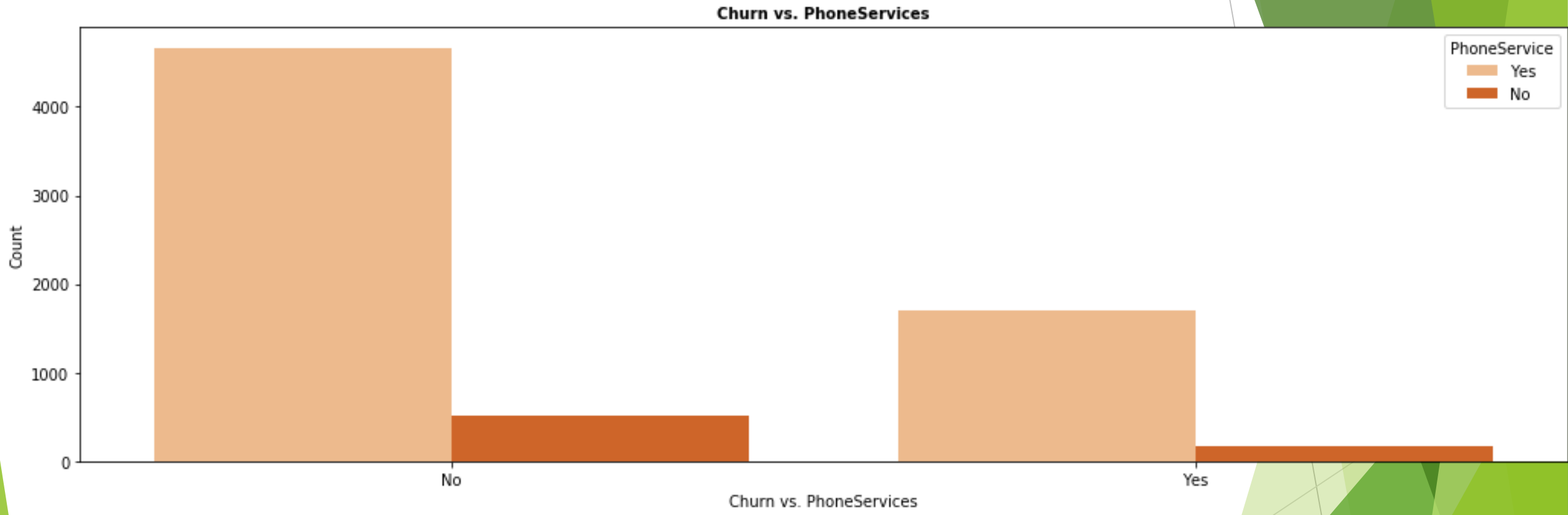
- Customers who paid $70/month were most likely to churn due to the expensive services offered. Those who paid less than $30 were satisfied with the services, which implicates they still recommended the service.

Churn vs. Total Charges



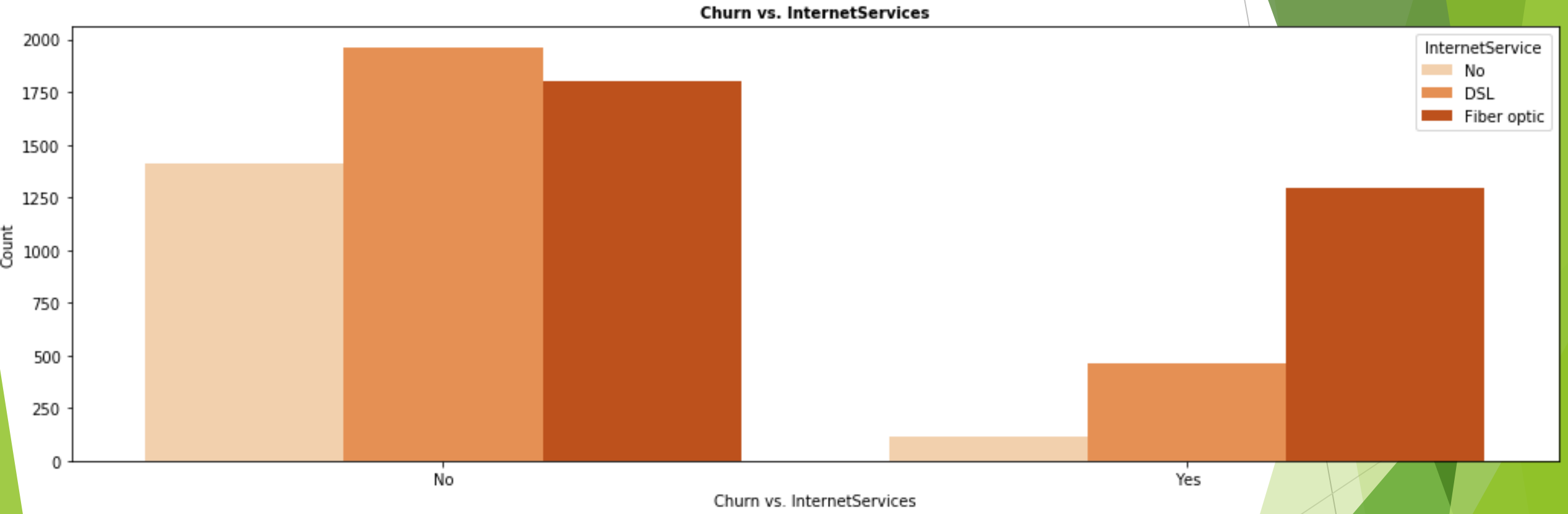
- Those churned customers are the ones that paid a total of less than $1,000 month. The more they paid in total, they less of a chance they would churn the company.

Churn vs. Phone Services



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

Churn vs. Internet Services



- Those churned customers were the ones who had Fiber Optic. Because Fiber Optic is faster than DSL, it is an implication that it was more expensive..

Unit 5 Specialization

Are senior citizens more likely to churn?

Are tenured customers more likely to churn?

Data - Are senior citizens more likely to churn?

Gender (Based on Non-Senior Citizen)
Female 0.239384
Male 0.232808
Name: Churned, dtype: float64
KstestResult(statistic=0.5920446653482425,
pvalue=0.3328551101416548 6)

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

Analysis - Are senior citizens more likely to churn?

Because both p-values were 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 connect well.

Data - Are tenured customers more likely to churn?

Gender (Tenured = 1)
Female 0.654930
Male 0.589666
Name: Churned, dtype: float64
KstestResult(statistic=0.7222925868319272, pvalue=0.1542428146570053 7)

Gender (Tenured = 2)
Female 0.523077
Male 0.509259
Name: Churned, dtype: float64
KstestResult(statistic=0.6947147446423894, pvalue=0.1863981742775230 6)

Gender (Tenured = 3)
Female 0.495050
Male 0.444444
Name: Churned, dtype: float64
KstestResult(statistic=0.6716393567181147, pvalue=0.2156414241129870 6)

Gender (Tenured = 4)
Female 0.534091
Male 0.409091
Name: Churned, dtype: float64
KstestResult(statistic=0.6587635262502591, pvalue=0.2328846620343151 8)

Because both p-values were higher than 0.05 for Tenure (How long a customer has been with the company) vs. Churning, we cannot reject the null hypothesis. Therefore, the relationship between tenure and churning do not connect well.

However, as the tenured of customers got younger, the p-value kept going lower.

Analysis - Are tenured customers more likely to churn?

Because both p-values were higher than 0.05 for Tenure (How long a customer has been with the company) vs. Churning, we cannot reject the null hypothesis. Therefore, the relationship between tenure and churning do not connect well.

However, as the tenured of customers got younger, the p-value kept going lower.

Prediction Models

Random Forest Classifier

Logistic Regression

Support Vector Machine

Gradient Booster Classifier

Random Forest Classifier

The Random Forest Classifier results are below:
Accuracy: 78.98722195929957
F1: 72.14916759683663
Precision: 73.34214844295491
Recall: 71.28588645950586
Cross Validation: [77.37226277 76.44552648 75.47169811]

Here is the confusion matrix below:
[[1358 185]
[259 311]]

Random Forest Classifier had the second lowest accuracy score of 78.99% and had the lowest overall for cross-validation scores.

Logistic Regression

The Logistic Regression results are below:
Accuracy: 79.08187411263606
F1: 71.66604777368418
Precision: 73.59812103906592
Recall: 70.46565701356438
Cross Validation: [79.07542579 79.0626902 77.90626902]

Here is the confusion matrix below:
[[1376 167]
[275 295]]

Logistic Regression had the second lowest accuracy score of 79.08%, but it was ranked the lowest for the cross-validation scores.

Support Vector Machine

The Support Vector Machine results are below:
Accuracy: 79.9810695693327
F1: 71.76141305566938
Precision: 75.49337570559982
Recall: 69.97504292162682
Cross Validation: [79.31873479 78.51491175 78.63664029]

Here is the confusion matrix below:
[[1415 128]
[295 275]]

Support Vector Machine had the second highest accuracy score with 79.98%, but it ranked first in cross-validation scores.

Gradient Booster Classifier

The Gradient Boosting Classifier results are below:
Accuracy: 81.16422148603881
F1: 73.93137011779294
Precision 77.00831847890672
Recall: 72.22334026901342
Cross Validation: [78.89294404 78.75836884 78.51491175]

Here is the confusion matrix below:
[[1414 129]
[269 301]]

Gradient Booster Classifier had the highest accuracy score with 81.16%, but it ranked 2nd in cross-validation scores.

Final Data Analysis (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 Logistic Regression because it was 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. 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.

Gradient Boosting, Random Forest, Support Vector Classifier, and Logistic Regression were 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.

Final Data Analysis (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?

For those models, the cross-validation scores were inconsistent with the four models that I used in terms of ranking. Despite this, the models work because of the consistent scores 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.

Final Data Analysis (Part 3)

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 was **Gradient Boosting**.

77.01% 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), 72.22% 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 was lower because there was more false negative observations in the dataset.

81.16% of **accuracy** means (True Positive + True Negative)/(True Positive + False Positive + False Negative + True Negative)). This was the highest out of the three because there was more true negative observations.

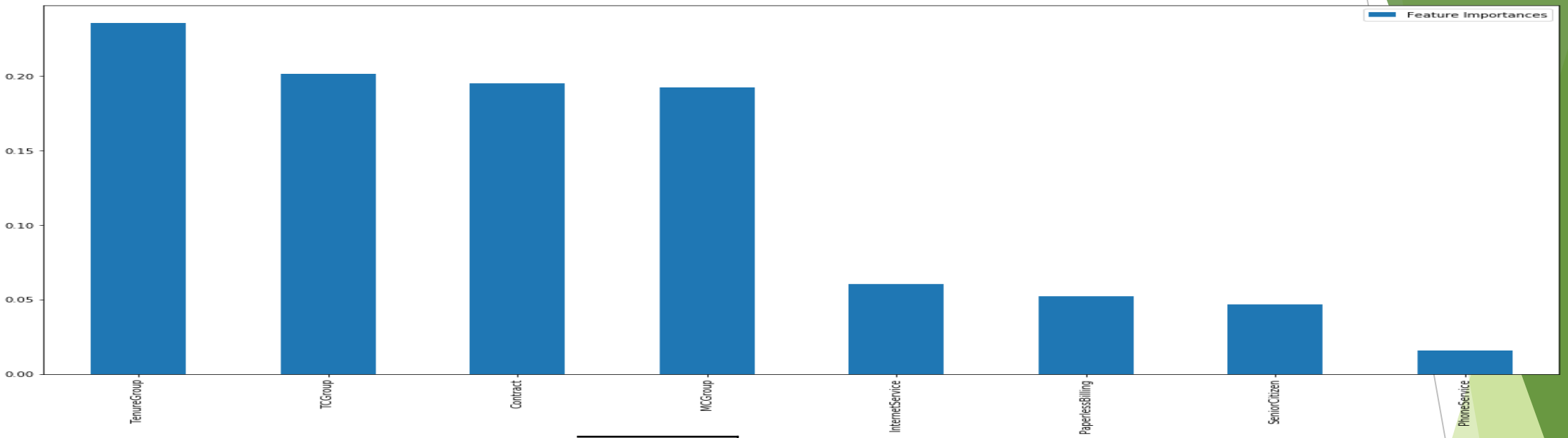
Final Data Analysis (Part 4)

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:

- Make, 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

Feature Importances



Feature Importances	Percentage
TenureGroup	0.235709
TCGroup	0.201333
Contract	0.195182
MCGroup	0.192471
InternetService	0.060306
PaperlessBilling	0.052348
SeniorCitizen	0.046894
PhoneService	0.015756

Conclusion

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.