Predicting customer churn at telecoms company, Telco

Data analysis and interpretation

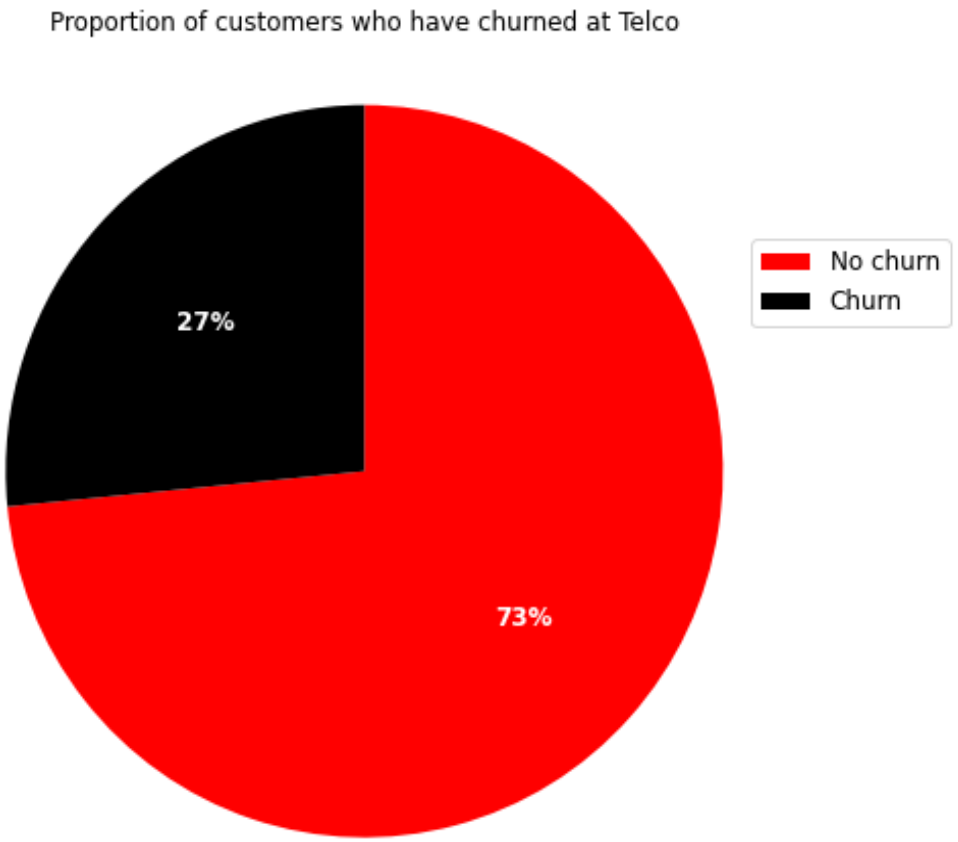
By Amy Birdee

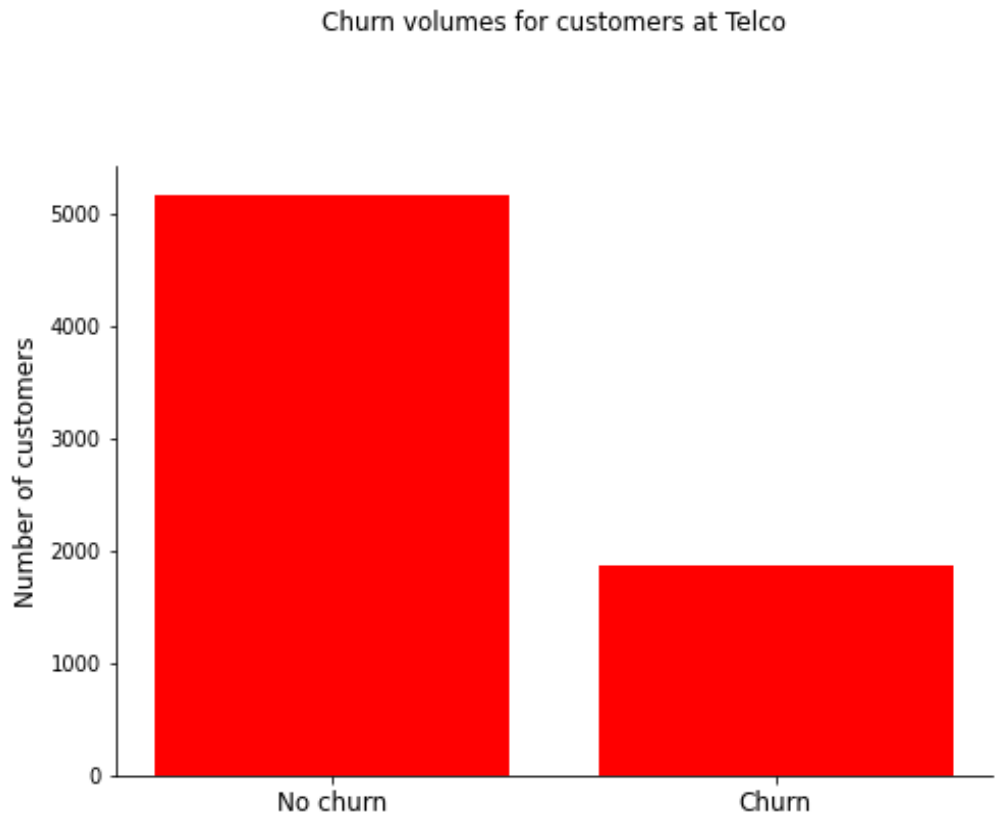
Introduction

- The data consist of details of circa 7,000 customers at a telecommunications company called Telco including whether or not they churned
- This project aims to segment the customer data and build a classifier model which will predict whether or not a customer will churn
- The main findings in the data have been presented in graphical format and the data analysis has been carried out in Python
- The models used include a logistic regression and a random forest classifier
- The Python libraries used include Pandas, Numpy, Matplotlib, Seaborn and Sklearn

How many customers churned at Telco?

Churn rates seem to be quite high at 27% which equates to 1,869 customers

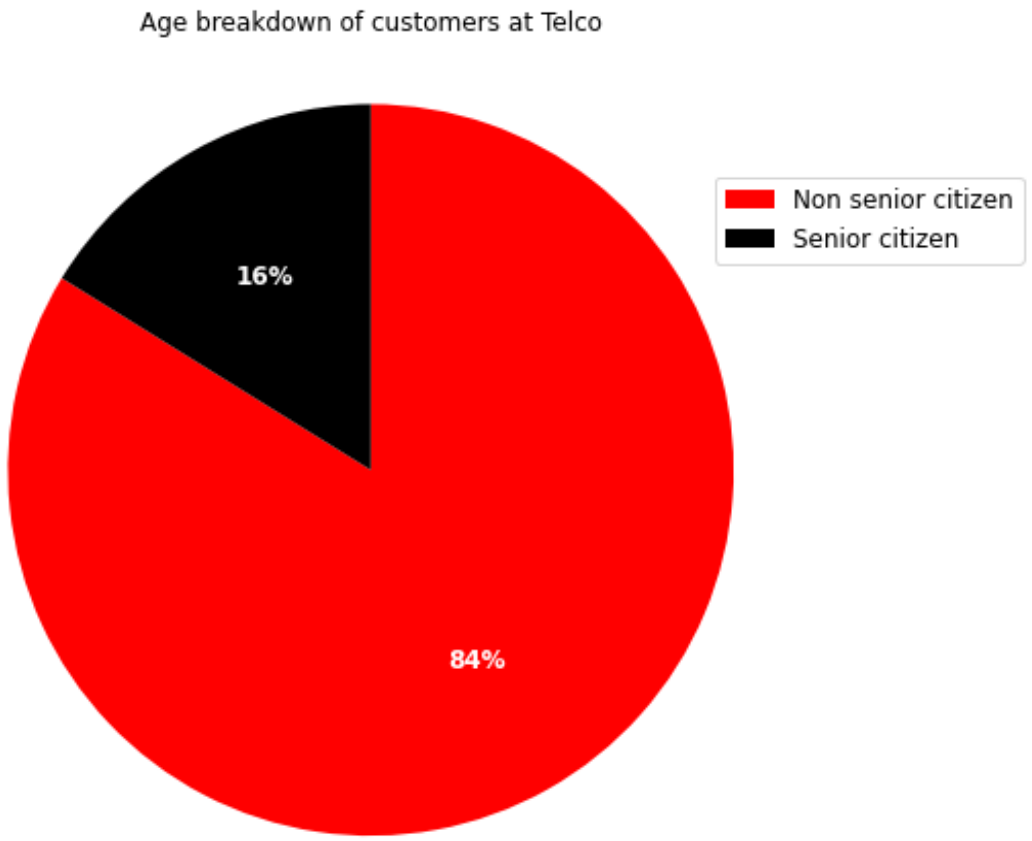


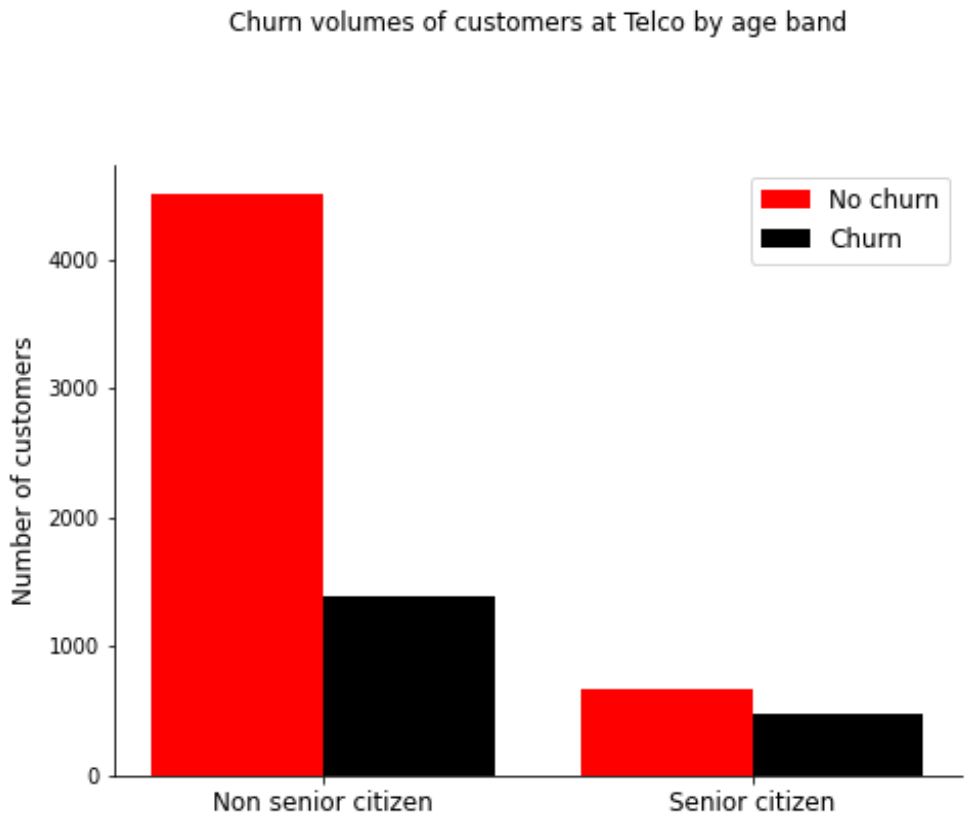


On a positive note, 5,174 customers did not churn over the same period

How does age profile affect churn at Telco?

Customers are broken down by senior citizen and non-senior citizen. The majority (84%) are non-senior citizens

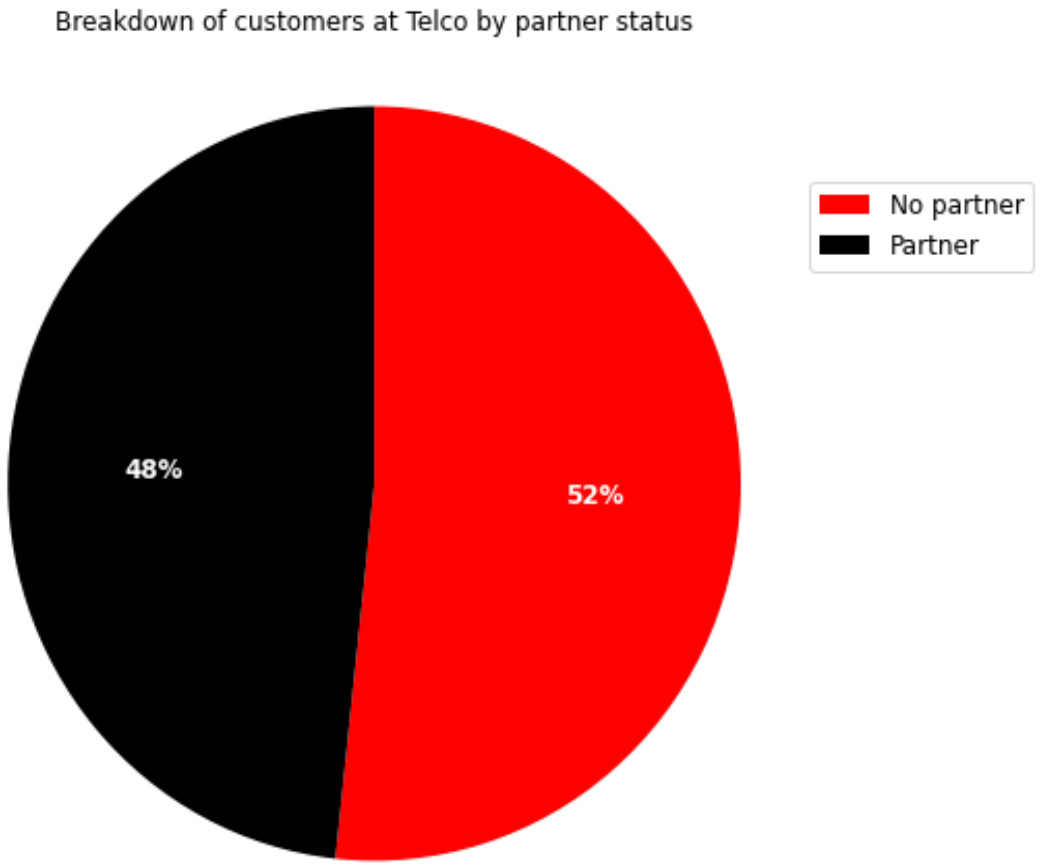


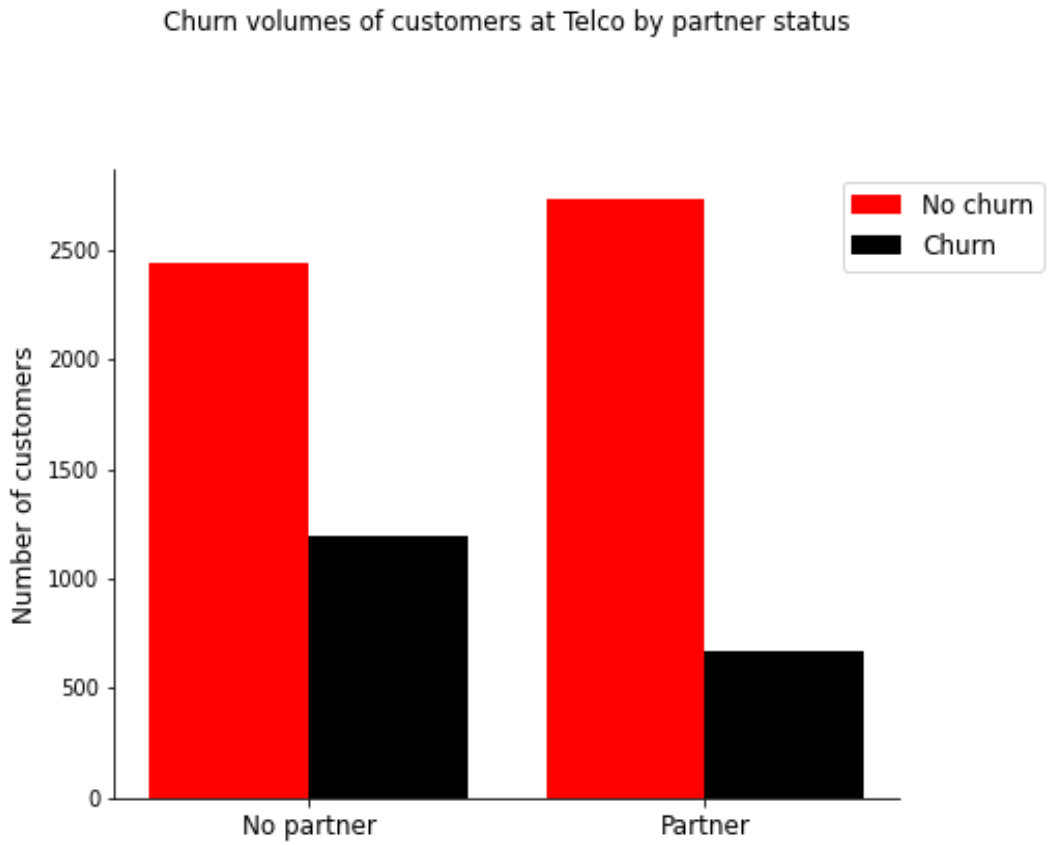


A higher number of non-senior citizens will churn compared to senior citizens but a higher *proportion* of senior citizens will churn – 42% of all senior citizens churned compared to 24% of all non-senior citizens

How does a customer’s partner status affect churn at Telco?

The split between having a partner and not having a partner is fairly even at 48% vs 52%

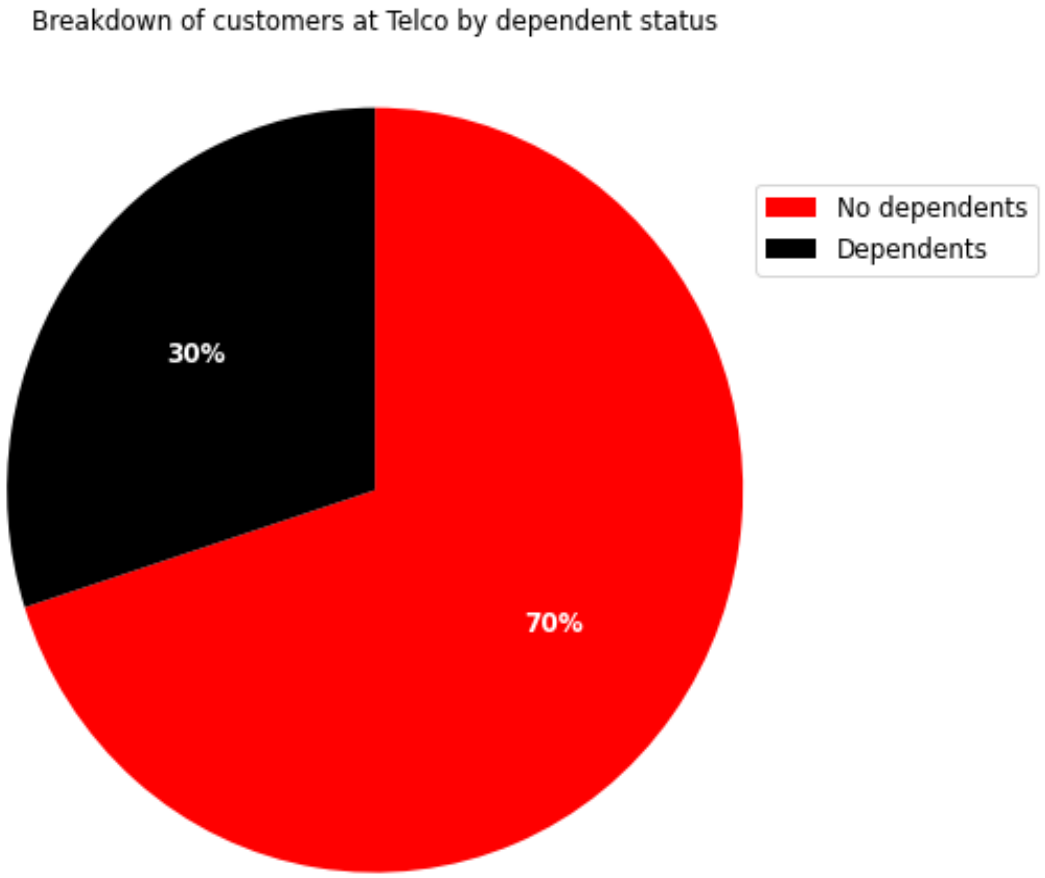


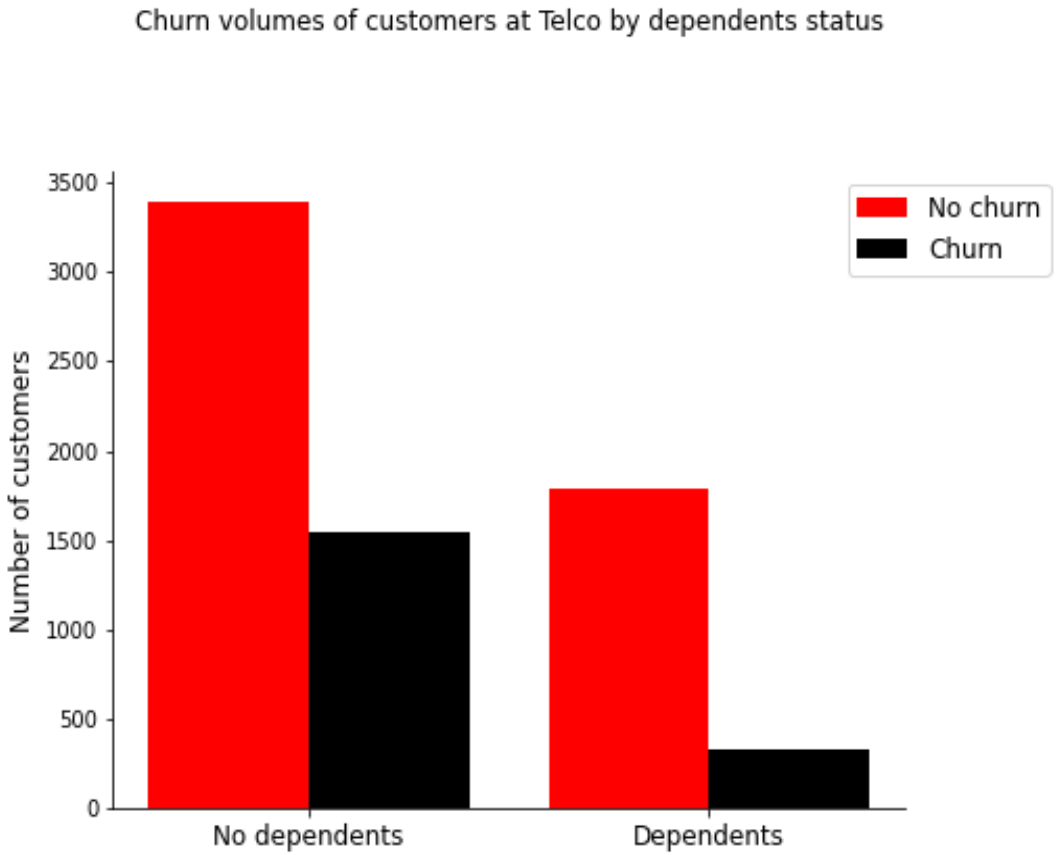


Those without a partner are more likely to churn with 531 more single customers churning compared with those that do have a partner

How does a customer's dependents status affect churn at Telco?

The majority of customers (70%) do not have any dependents and these are the customers most likely to churn

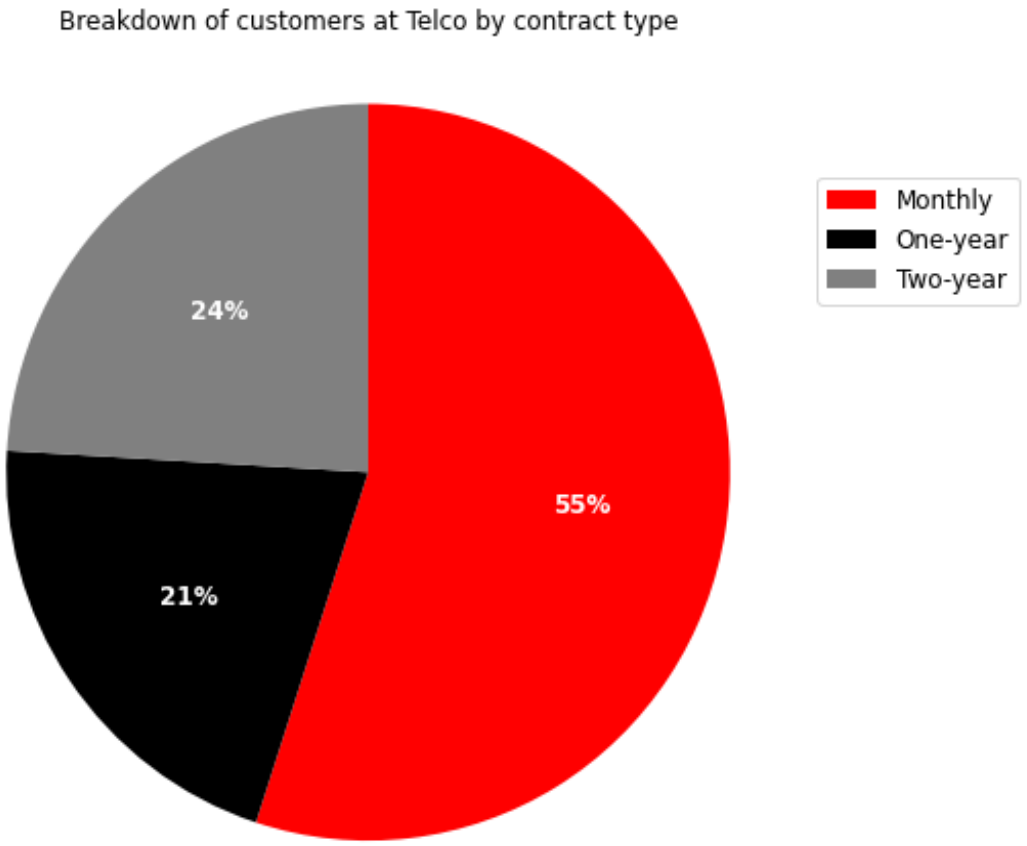


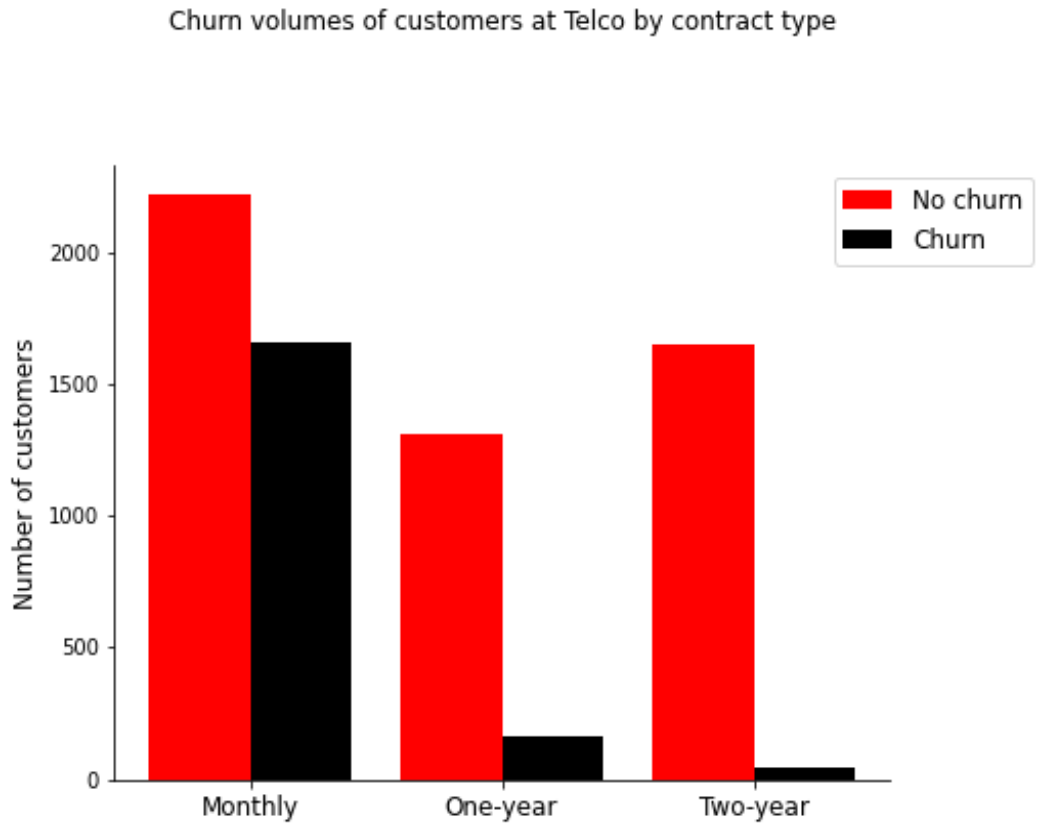


Out of all the customers without dependents, 31% churned compared to just 15% of those who do have dependents. This could be because those with dependents are too busy with caring duties to shop around for better deals.

How does the contract type affect churn at Telco?

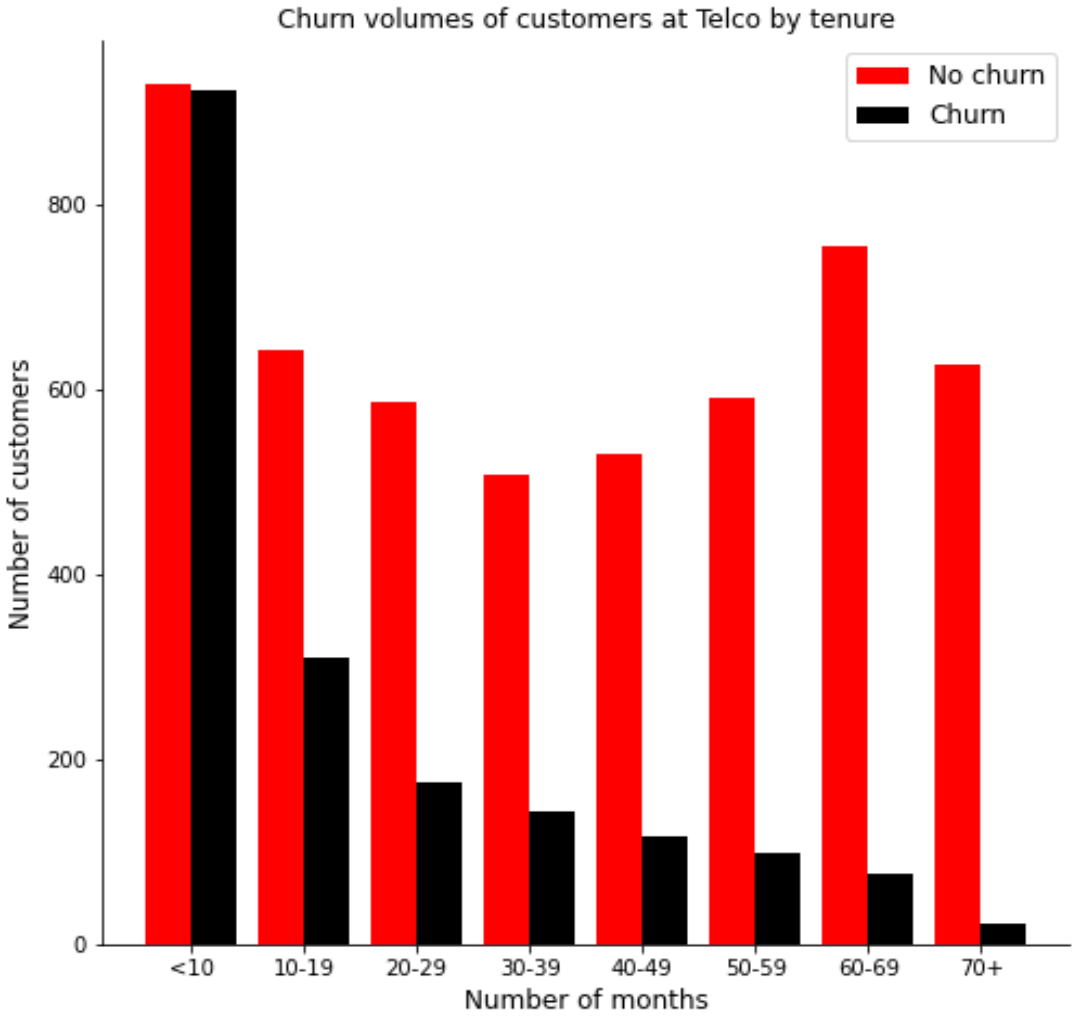
The majority of customers (55%) are on a monthly contract and this is also the group that churns the most





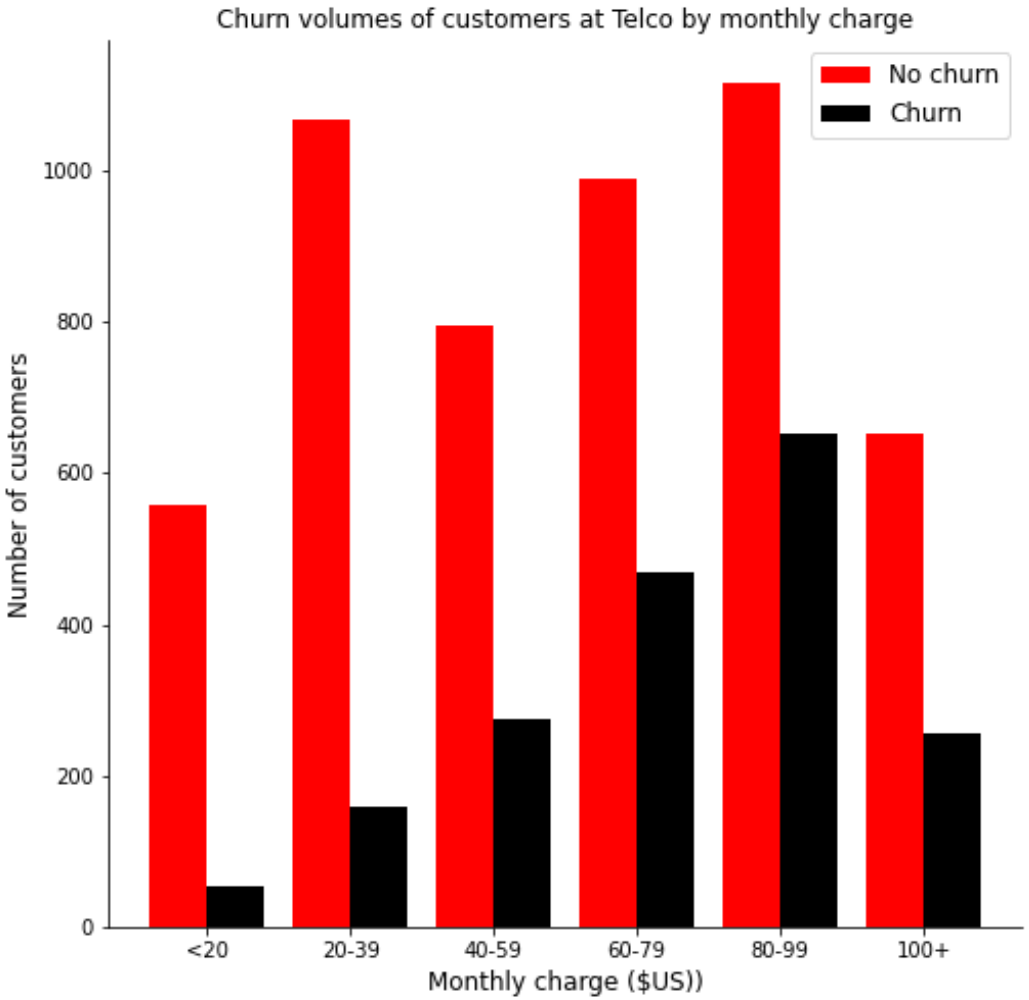
Those on one and two-year contracts are less likely to churn with 11% of those on one-year contracts churning and 3% of those on two-year contracts churning

How does the tenure affect churn at Telco?



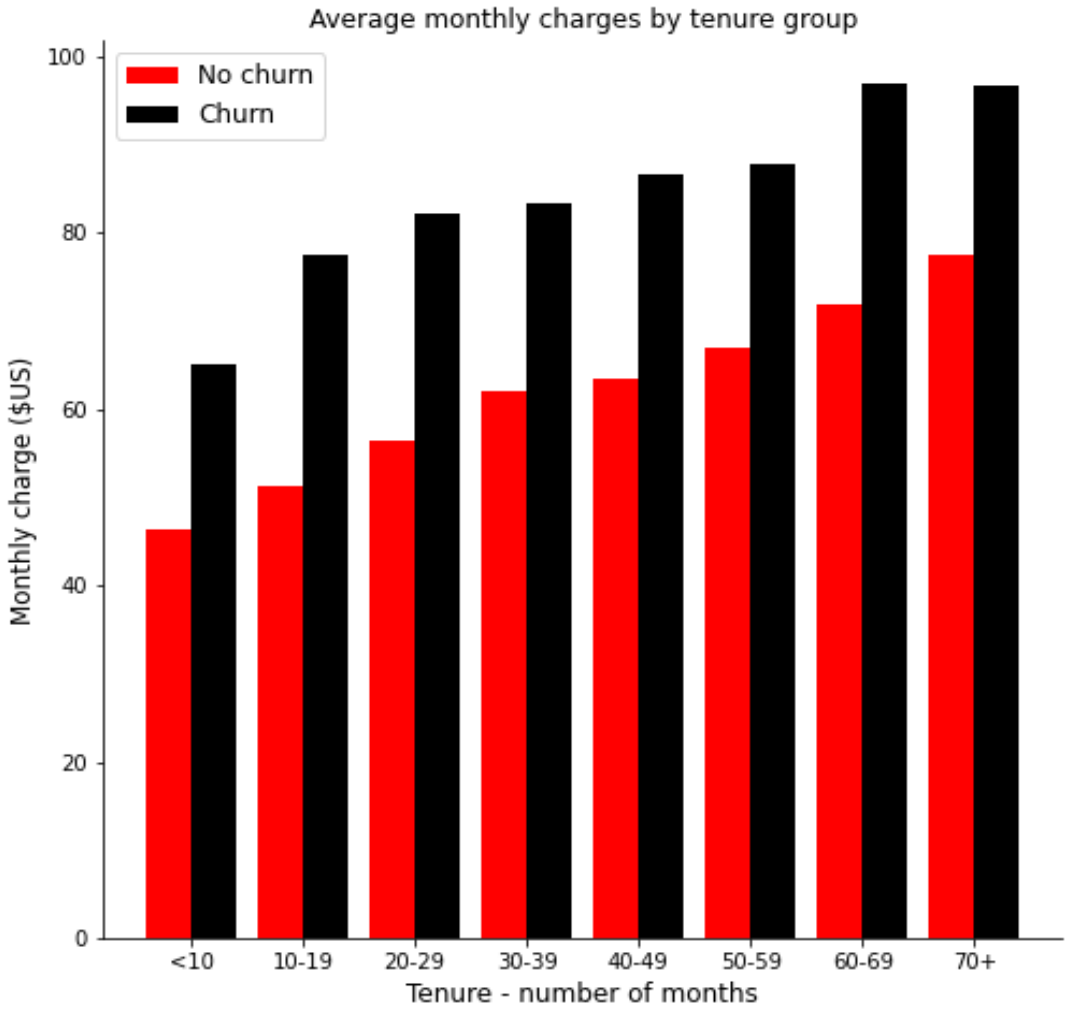
- Customers are far more likely to churn within the first ten months of joining Telco (this may be within a cooling off period)
- 50% of customers who join Telco churn within the first ten months
- Churn rates decrease rapidly after the first ten months and then continues to fall as customer loyalty builds

How does the monthly charge affect churn at Telco?



- Churn levels increase as the monthly charge increases – those paying less than $20 are least likely to churn
- Churn volumes are lower for those paying over $100 but this is likely because overall volumes are also lower for this category
- In percentage terms, churn rates for the $100+ group are quite high at 28% albeit still lower than for the $80-$99 group where churn rates are 37%

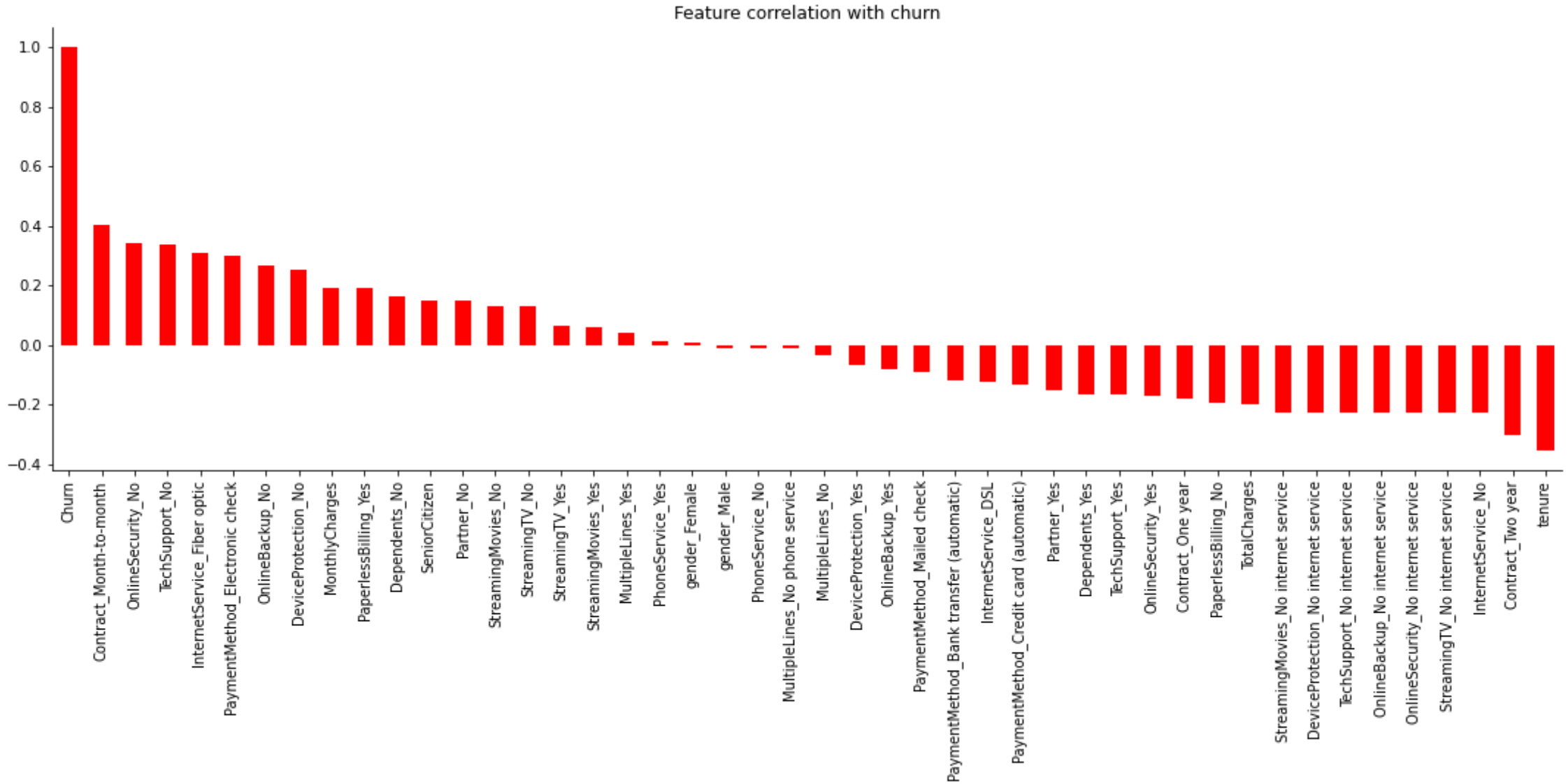
What is the relationship between tenure, monthly charge and churn?



- Customers paying a higher monthly charge are more likely to churn and this is true in all tenure groups
- Churn is more likely when customers are paying more than $70

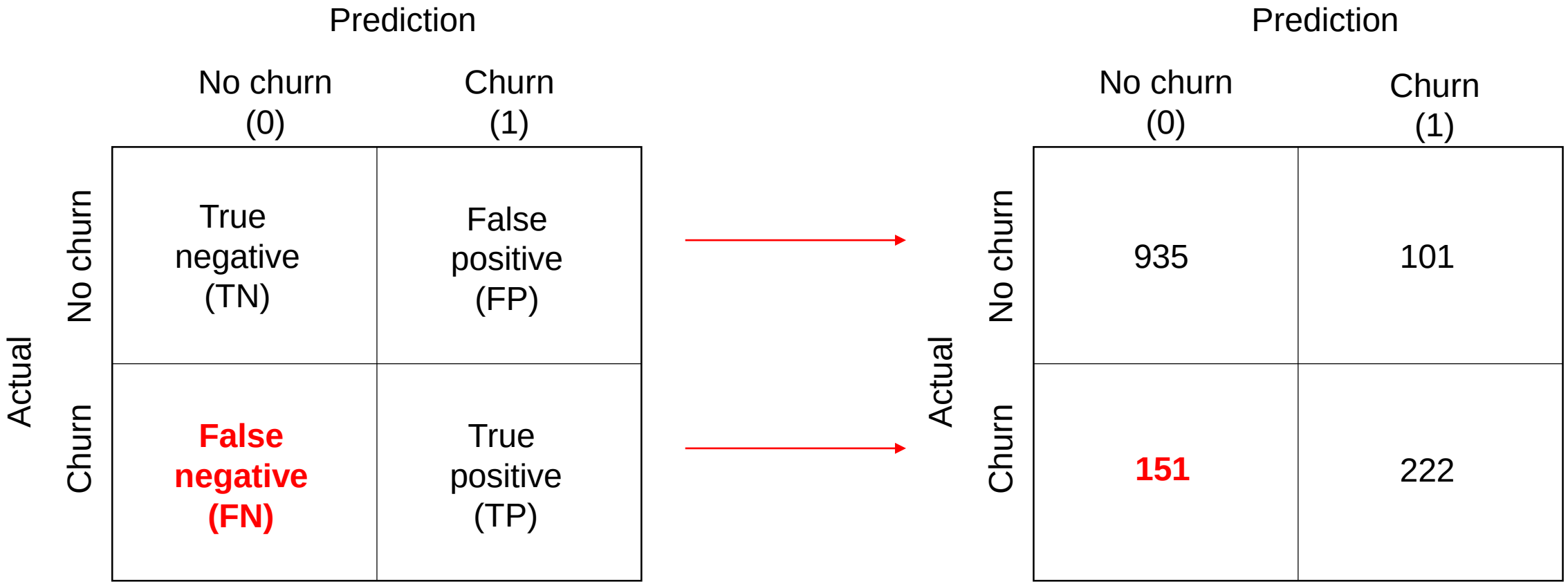
How do all features of the dataset correlate with churn?

Monthly contracts and no online security are most positively correlated with churn whereas tenure and two-year contracts are most negatively correlated



Evaluating logistic regression model: confusion matrix

- The aim of the analysis is to reduce the number of false negatives, i.e. those customers we predict will not churn but they do churn
- The logistic regression model classified 151 customers as false negatives out of the 1,409 included in the test dataset

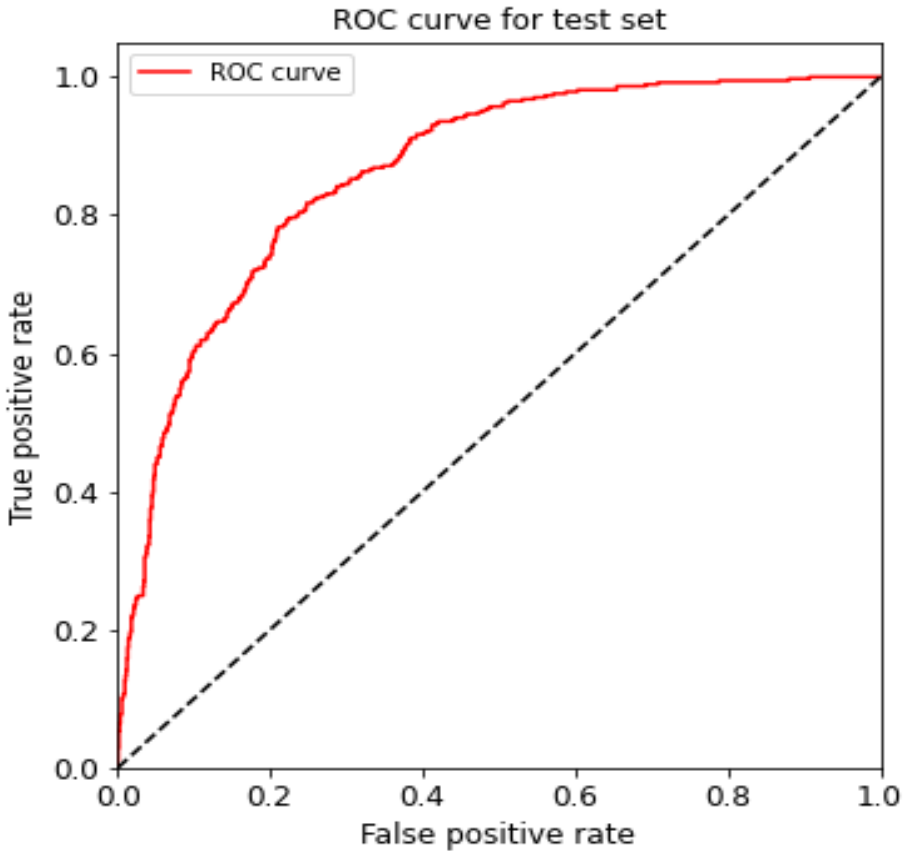


Evaluating logistic regression model: key metrics

- In a churn analysis, we want to reduce the instances of false negatives. Therefore, we are trying to optimise the recall metric which shows what percentage of the class we are interested in was captured by the model
- In other words, out of the customers that churned, what percentage did the model predict as ‘going to churn’?
- The recall score for the logistic regression model is 60%

	Precision	Recall	Accuracy	F1 score
Calculation	TP / (TP + FP)	TP / (TP + FN)	(TP + TN) / (TP + FP + TN + FN)	2 * (Precision * Recall) / (Precision + Recall)
Result	0.69	**0.60**	0.82	0.64

Evaluating logistic regression model: ROC curve and AUC

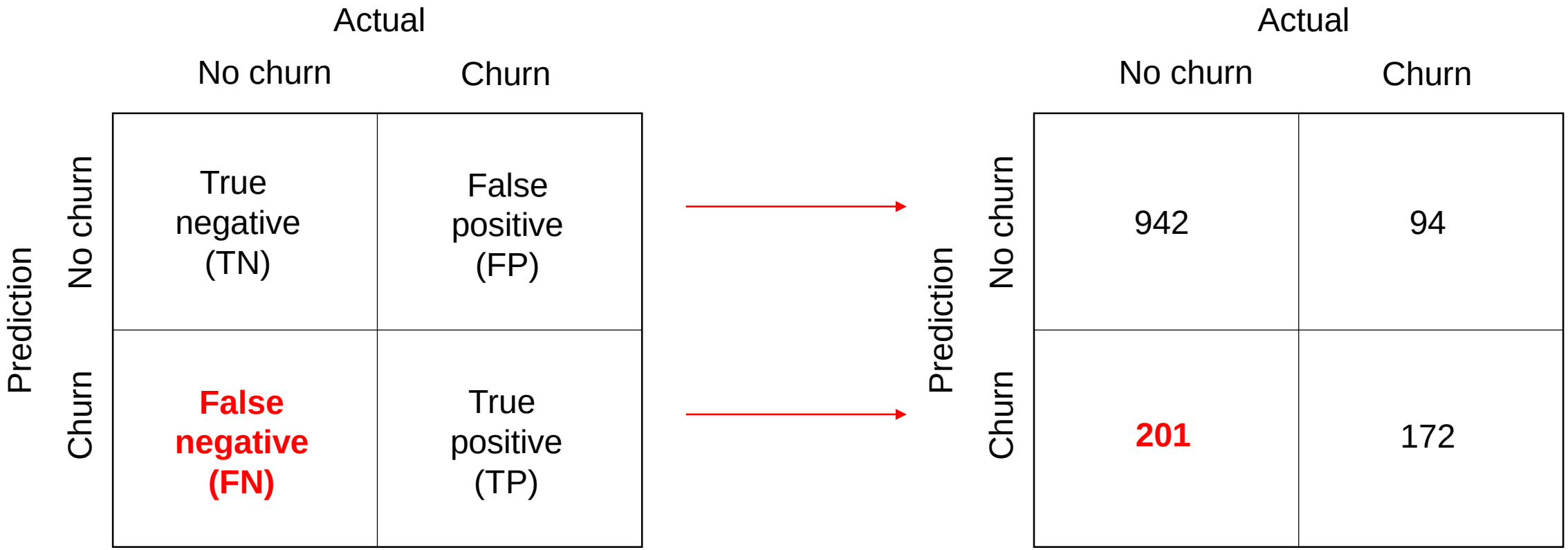


- The ROC (receiver operator characteristic) curve is in the upper portion of the grid and also quite far off the 50% line which is encouraging and suggests the model is fairly robust

- The AUC (area under curve) score is 0.86 which is also a pretty good score (AUC score can be between 0 and 1)

Evaluating random forest model: confusion matrix

The random forest model incorrectly classified 33% more false negatives than the logistic regression model suggesting that the random forest model requires more tuning before it can accurately fit the data



Evaluating random forest model: key metrics

- The random forest model performed worse than the logistic regression model across all four metrics
- The model resulted in a recall score of just 0.46 compared with 0.60 in the logistic regression model

	Precision	Recall	Accuracy	F1 score
Calculation	TP / (TP + FP)	TP / (TP + FN)	(TP + TN) / (TP + FP + TN + FN)	2 * (Precision * Recall) / (Precision + Recall)
Result	0.65	**0.46**	0.79	0.54

Conclusion

- In terms of customer demographics, customers who are younger and without a partner or dependents are more likely to churn
- Customers on monthly contracts are much more likely to churn than those on longer term contracts – 43% of customers on monthly contracts churned during the period
- Customers are more likely to churn within the first ten months of joining Telco. Churn rates fall considerably after 10 months and continue to decrease as customer loyalty builds
- Churn volumes increase as the monthly charge increases. Those paying less than $20 a month are least likely to churn and those paying more than $70 are most likely to churn
- The logistic regression model outperformed the random forest model in predicting churn volumes with higher precision, recall, accuracy and F1 scores
- The analysis could be improved by further tuning the random forest model to see if improved results are achieved. Model tuning that has taken place so far consists of amending the number of trees in the forest

Thank you

Contact details:

Amy Birdee

amybirdee@gmail.com