

# Churn Prediction

Quissuiven Tai



# Table of Contents

## 1. Exploratory Data Analysis and Feature Engineering

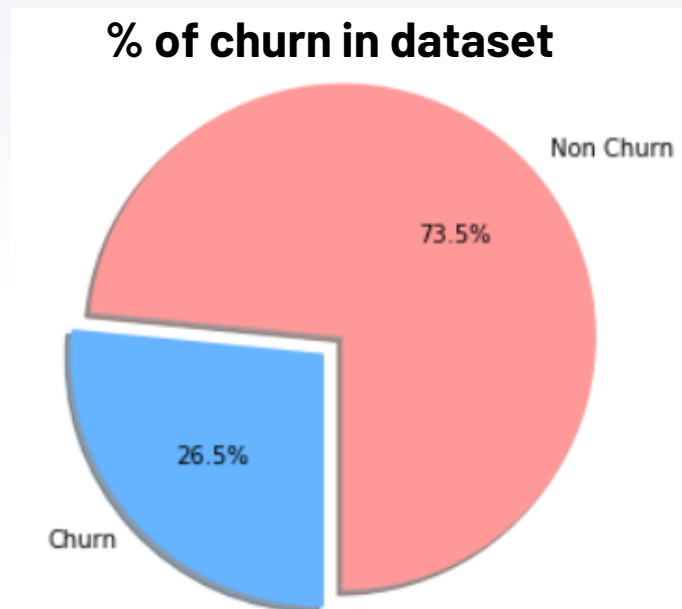
- ▶ Exploring the dataset
- ▶ Exploring individual features
- ▶ Exploring feature relations

## 2. Model Building and Feature Selection

- ▶ Baseline Model
- ▶ Resampling with Replacement
- ▶ Feature Selection
- ▶ Obtaining the Best Model
- ▶ Hyperparameter Tuning
- ▶ Most Important Features

## 3. Deep dives and Recommendations

# Exploring the dataset



- Unbalanced Dataset
- Model trained on the data may make predictions in favour of the majority class
- Can be handled by using a different performance metric such as AUROC score instead of accuracy

# Exploring Individual Features



## **Categorical variables**

(eg. Gender, InternetService)

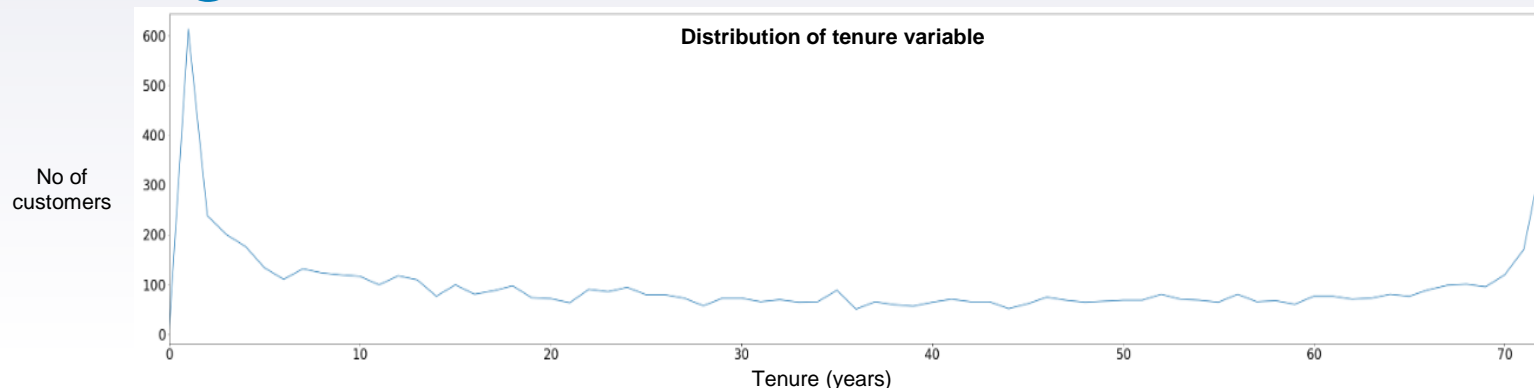
# 123

## **Numeric variables**

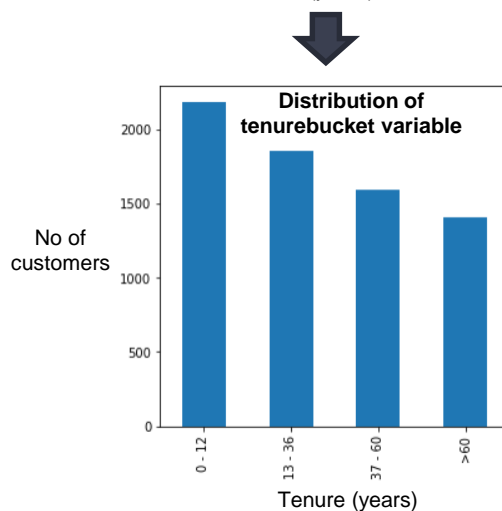
(eg. Tenure, MonthlyCharges)

- Have to be converted to numeric form to be used in models. For quick exploratory analysis, categorical variables are label encoded.
- Online Service variables such as OnlineSecurity, have a value called “No internet service”, which may be a repetition of InternetService variable. We will decide whether to convert the value to “No” or keep it, after exploring feature relations later.
- MultipleLines seems to be a breakdown of PhoneService. We shall explore their relationship and remove one if they are dependent later.
- Should visualize distributions to see if there are any patterns in the data.

# Exploring Numeric variables: tenure



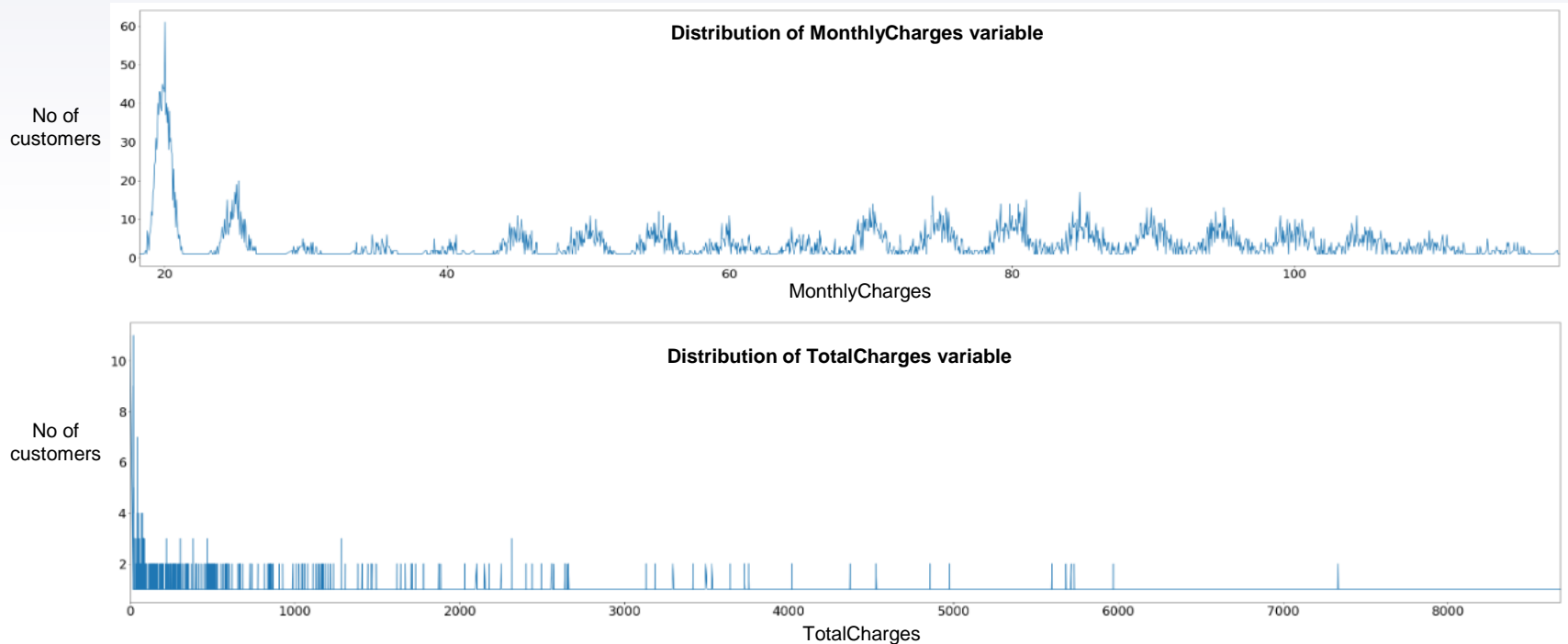
- From the above graph, there seems to be a large number of customers who have the shortest or longest tenures.
- Thus, it might be more representative to convert tenure into a categorical variable and split it into buckets.



The buckets of the tenurebucket variable are chosen such that

- the patterns in the shortest and longest tenures are captured
- each bucket has relatively equal number of observations

# Exploring Numeric variables: MonthlyCharges and TotalChages



Both graphs appear to have a similar pattern. Perhaps, MonthlyCharges and TotalCharges are highly correlated. We shall verify the relationship between the two variables in the next section.

# Pearson correlation is used to compare numeric variables; Chi squared for categorical

123 ↔ 123

Numeric variable 1

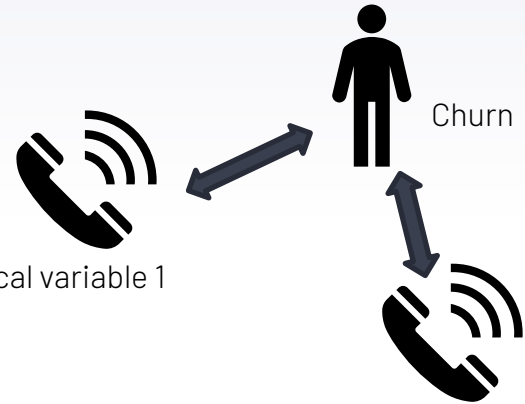
Numeric variable 2



Categorical variable 1

Categorical variable 2

If two features are highly correlated/dependent with each other, we will remove one as they have similar representations.

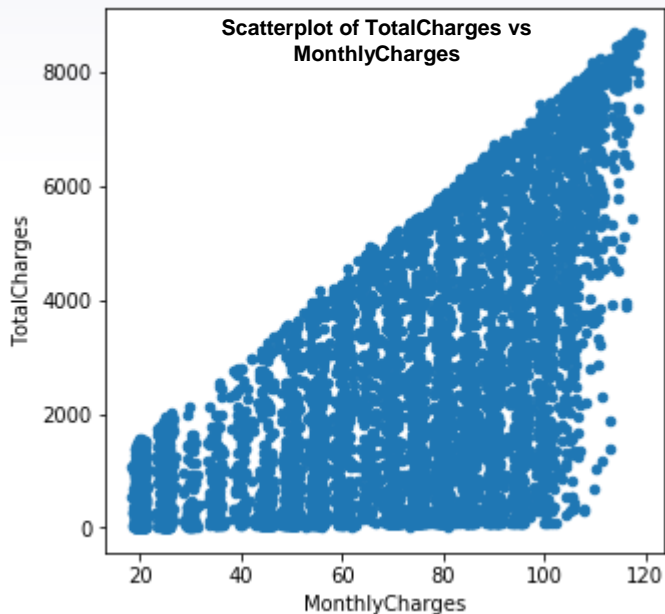


Categorical variable 1

Categorical variable 2

If one feature is more dependent with the Target Variable than another feature, and both are similar features, we shall pick the feature that is more dependent.

# Exploring relationship between MonthlyCharges and TotalCharges



Correlation = 0.65

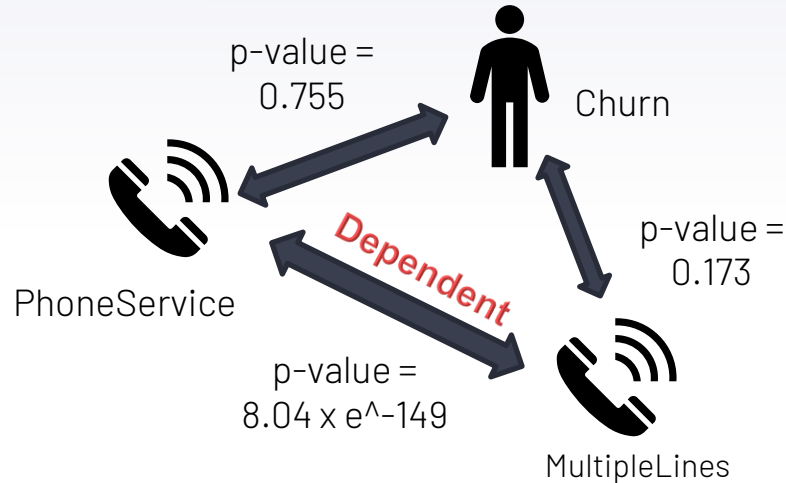
123 ↔ 123

MonthlyCharges      TotalCharges

- MonthlyCharges is highly correlated with TotalCharges.
- As both have the same representation and MonthlyCharges is more granular, we will only include MonthlyCharges in the model.



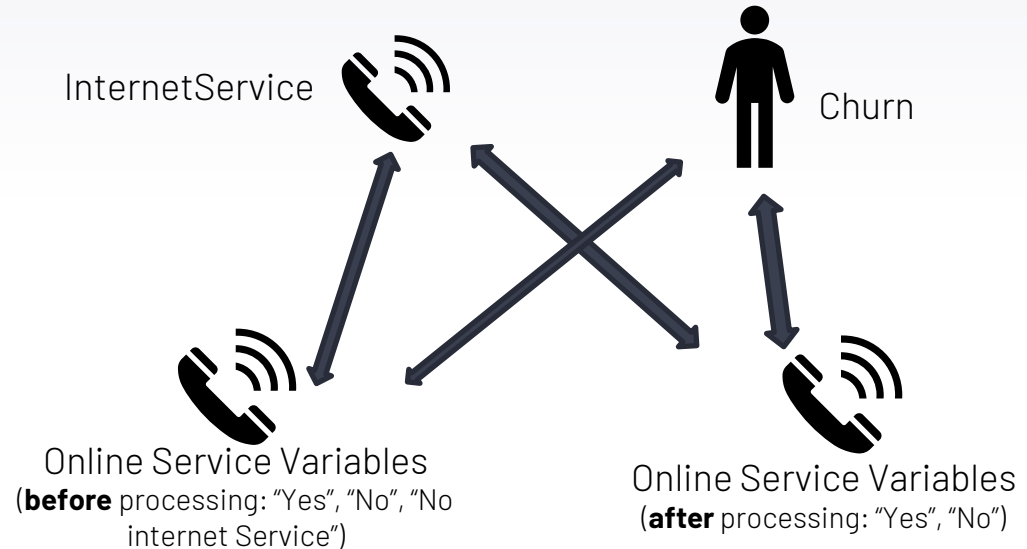
# Exploring relationship between PhoneService and MultipleLines and Churn



- p-value =  $8.04 \times 10^{-149}$  is less than 0.05, therefore PhoneService and MultipleLines are dependent. One of them can be excluded from the model.
- 0.173 is less than 0.755, therefore MultipleLines is relatively less independent with Churn than PhoneService. Thus, MultipleLines will be included while PhoneService will be excluded from the model.

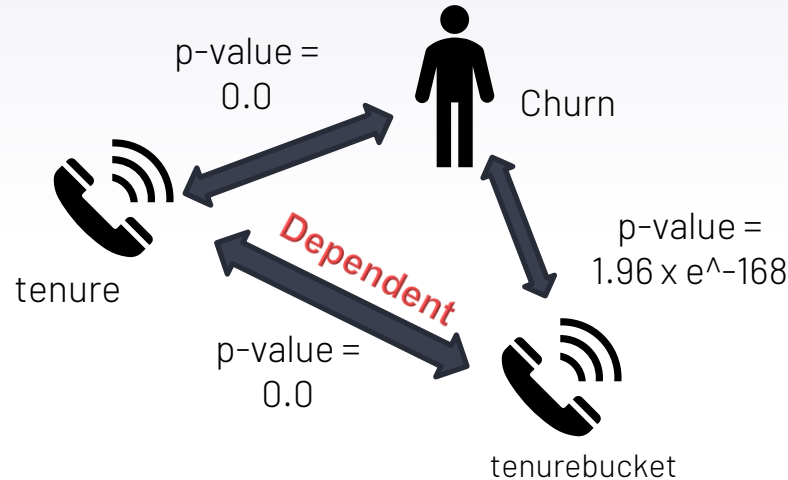
# Exploring relationship between Online Service Variables and InternetService and Churn

- Online Service variables, before and after processing, are dependent with InternetService and Churn.
- However, Online Service variables before processing are more dependent with InternetService (p-value = 0.0), meaning that they have very similar representations with InternetService.
- Thus, we will keep Online Service variables after processing in the model instead.



\*Online Service variables include "OnlineSecurity", "OnlineBackup", "DeviceProtection", "TechSupport", "StreamingTV", "StreamingMovies"

# Exploring relationship between tenure and tenurebucket and Churn



- As tenurebucket was derived from tenure, both variables are very dependent with each other.
- However, tenure is more dependent with Churn than tenurebucket. Thus, tenure will be included and tenurebucket will be excluded from the model.

# Building the Baseline Model

## Features used:

All the features in original dataset except customerID,  
TotalCharges and PhoneService

- LabelEncoded all categorical variables
- Converted "No internet service" to "No"

## Train Test Split:

0.7/0.3

(higher than standard 0.2 to prevent overfitting since dataset is small)

## Model used:

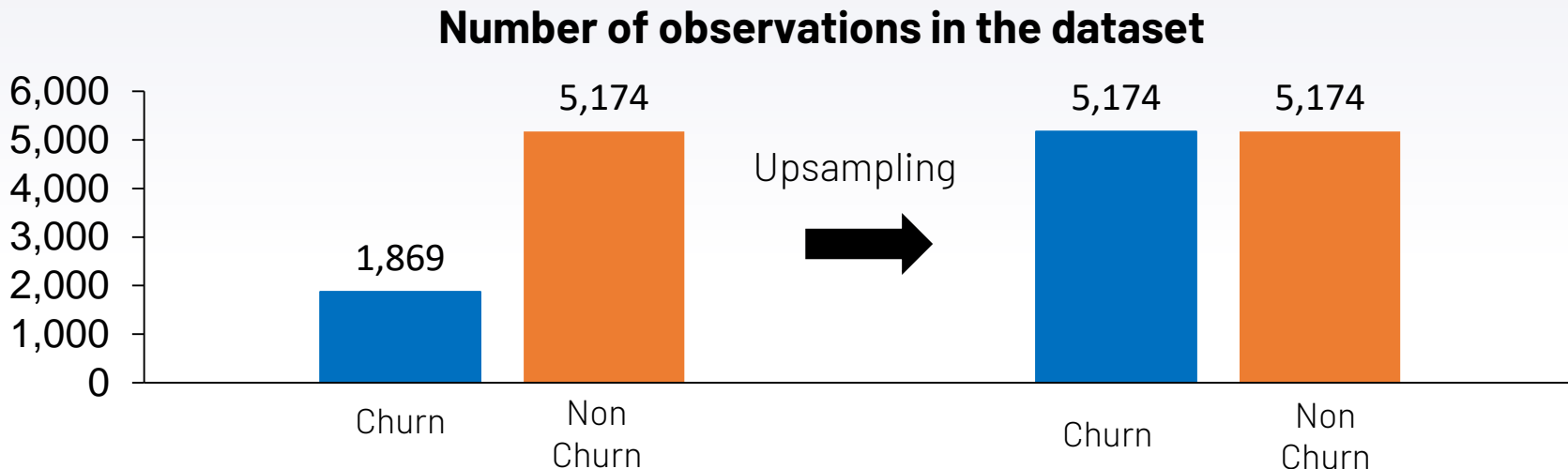
Logistic Regression

## AUROC Score:

0.839

Model serves as a baseline where performance can be improved by resampling with replacement and feature selection.

# Resampling with Replacement yielded higher AUROC score



	Unbalanced Dataset	Balanced Dataset
AUROC score (Baseline model)	0.839	0.849

As fitting the baseline model on the balanced dataset gives a higher AUROC score, the balanced dataset will be used subsequently, instead of the unbalanced dataset.

## Tree-Based Feature Importance was used for Feature Selection due to higher AUROC score

	Recursive Feature Elimination	Tree-Based Feature Importance
AUROC score	0.844	0.849
No of features	16	17

Following the Tree-Based method, all the features have feature importance of greater than 0.01. Thus, all the features will be included in the model.

# Before fitting other models, the features were preprocessed

123

Numeric variables  
(eg. Tenure, MonthlyCharges)  
were preprocessed  
By Min – Max Scaler

- Algorithms such as K-NN are sensitive to magnitude



Categorical variables  
(eg. Gender, InternetService)  
were preprocessed by  
One Hot Encoder

- As Label Encoding may cause some algorithms to misinterpret the values

# The Top 3 models were tree-based models, with the best model being Random Forest

Model	AUROC score
Random Forest	0.951
CatBoost	0.881
Gradient Boosting Classifier	0.868

This might be because tree-based models are able to capture non-linearities in the data that other models can't. It is also interesting to note that gradient boosting does not perform as well as the RF model, perhaps due to noise in the data.



# Grid Search was conducted to tune the hyperparameters of the RF model



**Best parameters:**      `max_depth = 25,    n_estimators = 800,    min_samples_leaf = 1`

**Final AUROC score:**      0.964

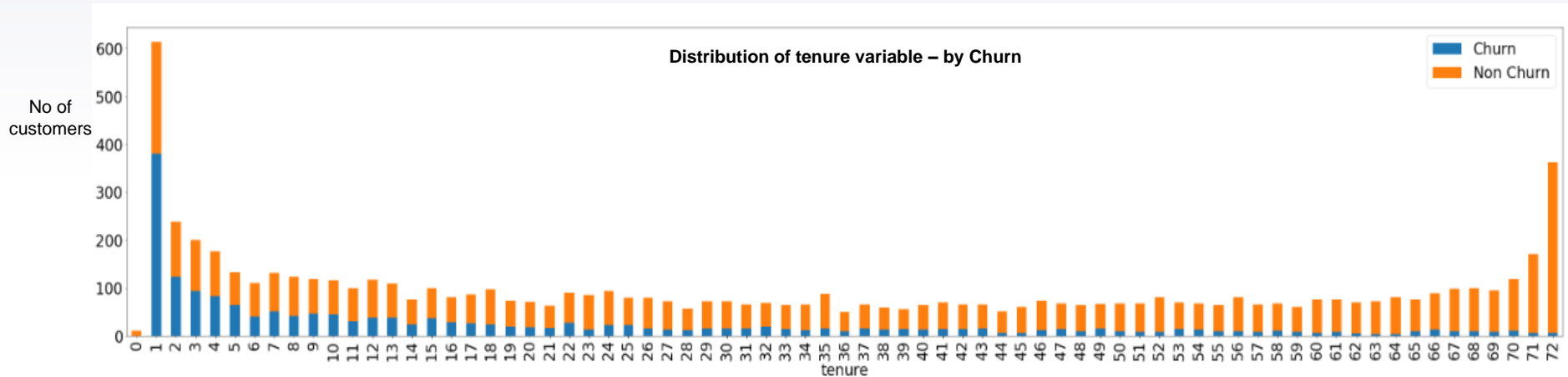
With the best parameters, a new RandomForestClassifier was trained and can be used to identify the most important features.



# Most Important Features

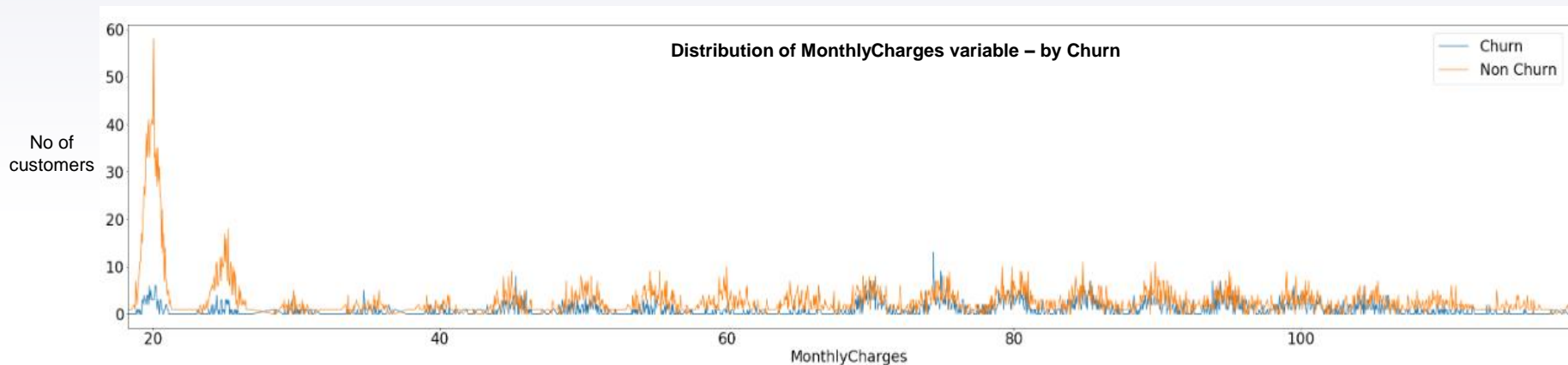
Feature	Feature importance
tenure 	0.1864
MonthlyCharges	0.1845
Contract (Month-to-Month)	0.0886
Contract (Two year) 	0.0397
InternetService (Fiber optic)	0.0375
PaymentMethod (Electronic check)	0.0328

# Tenure – Deep dive and Recommendation



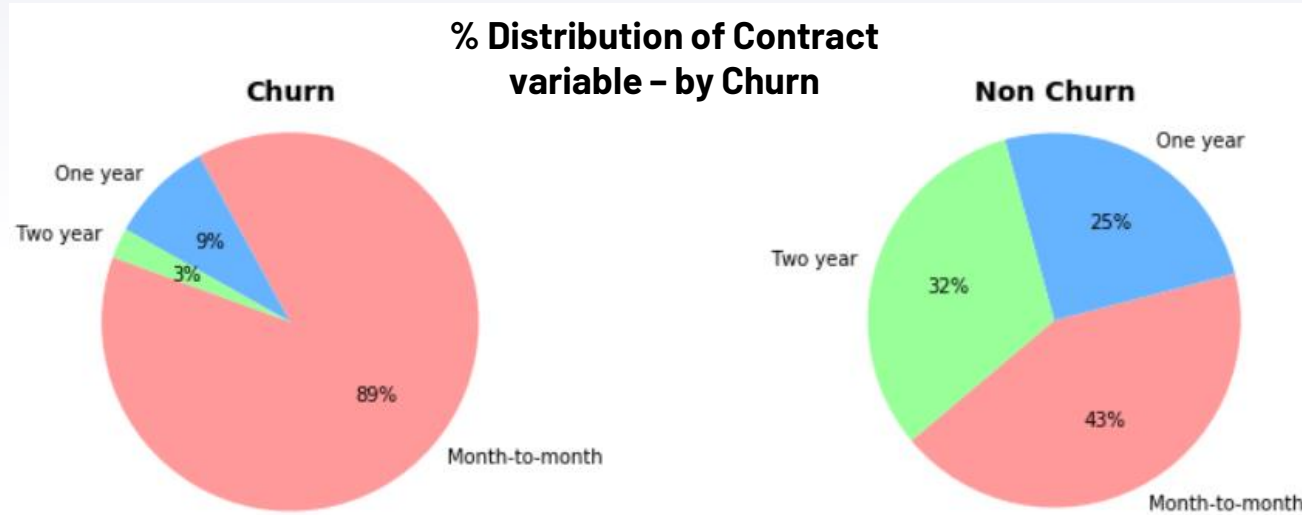
- Most people seem to churn at the start, especially after the first year.
- **Recommendation:** To overcome this problem, the company can come up with long term contracts such as 5 years to lock-in customers such that they are less likely to churn.

# MonthlyCharges – Deep dive and Recommendation



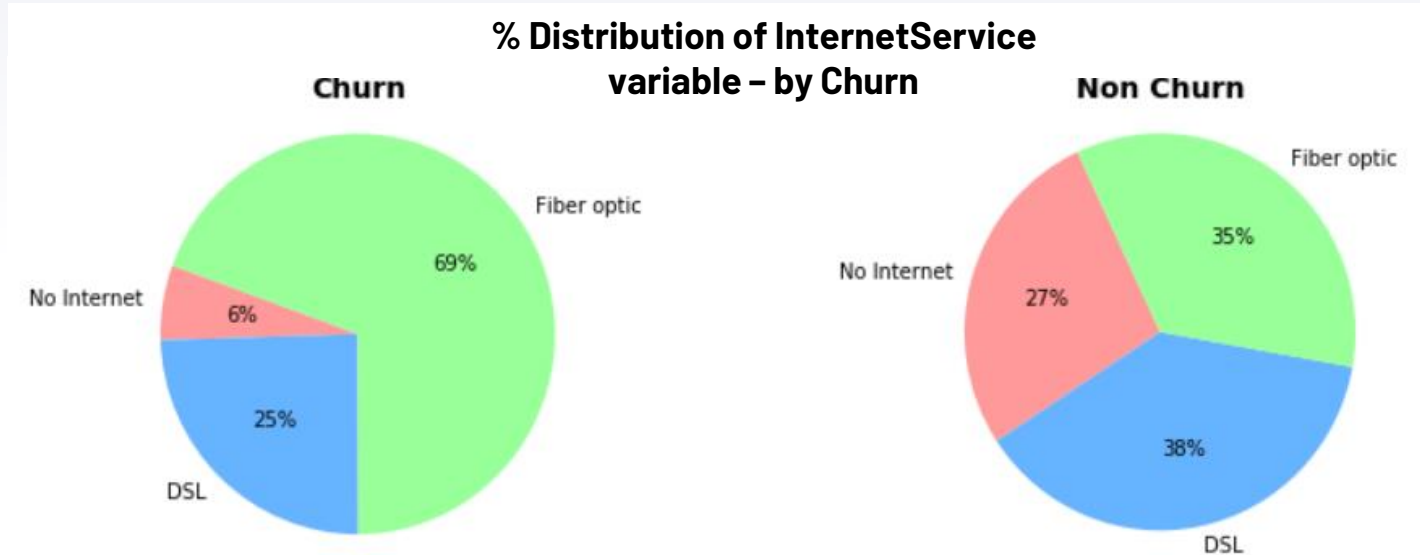
- Churn remains relatively low across the range of monthly charges. However, there is a large number of customers retained when monthly charge is around 20 or 25.
- **Recommendation:** To better retain customers, the company should strive to keep monthly charges to 20 or 25.

# Contract – Deep dive and Recommendation



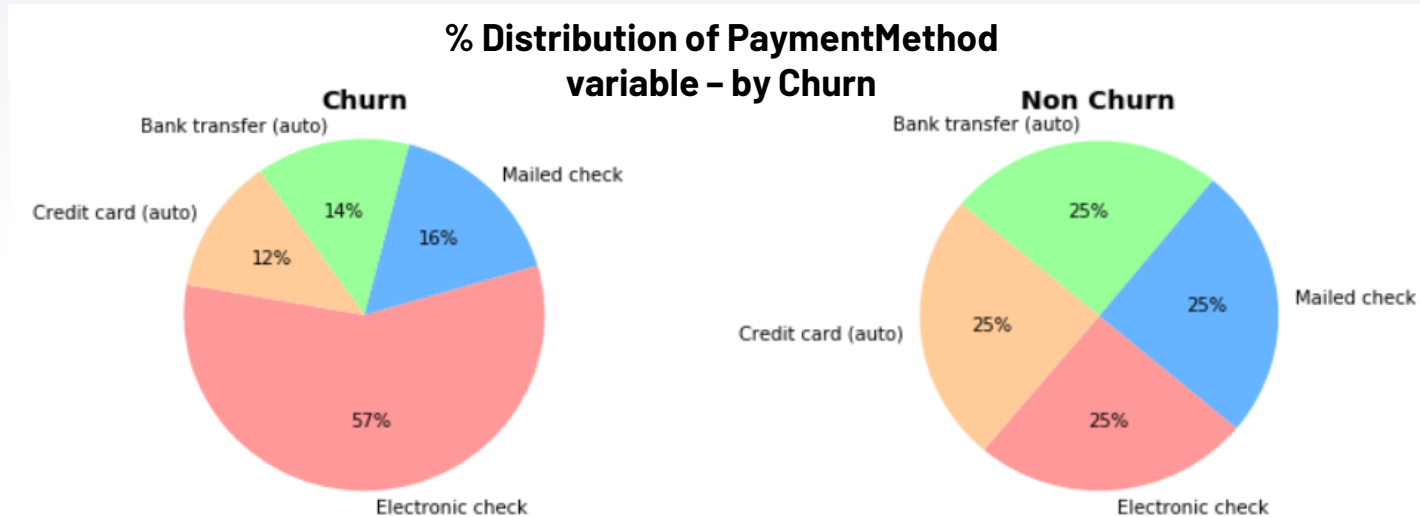
- 89% of customers who churned have month-to-month contracts. When contracts are so short, it is easy for customers to switch to other alternatives or competitors.
- **Recommendation:** To remedy this, the company can remove or reduce their offerings of month-to-month contracts and give longer term contracts, preventing customers from switching so easily.

# InternetService – Deep dive and Recommendation



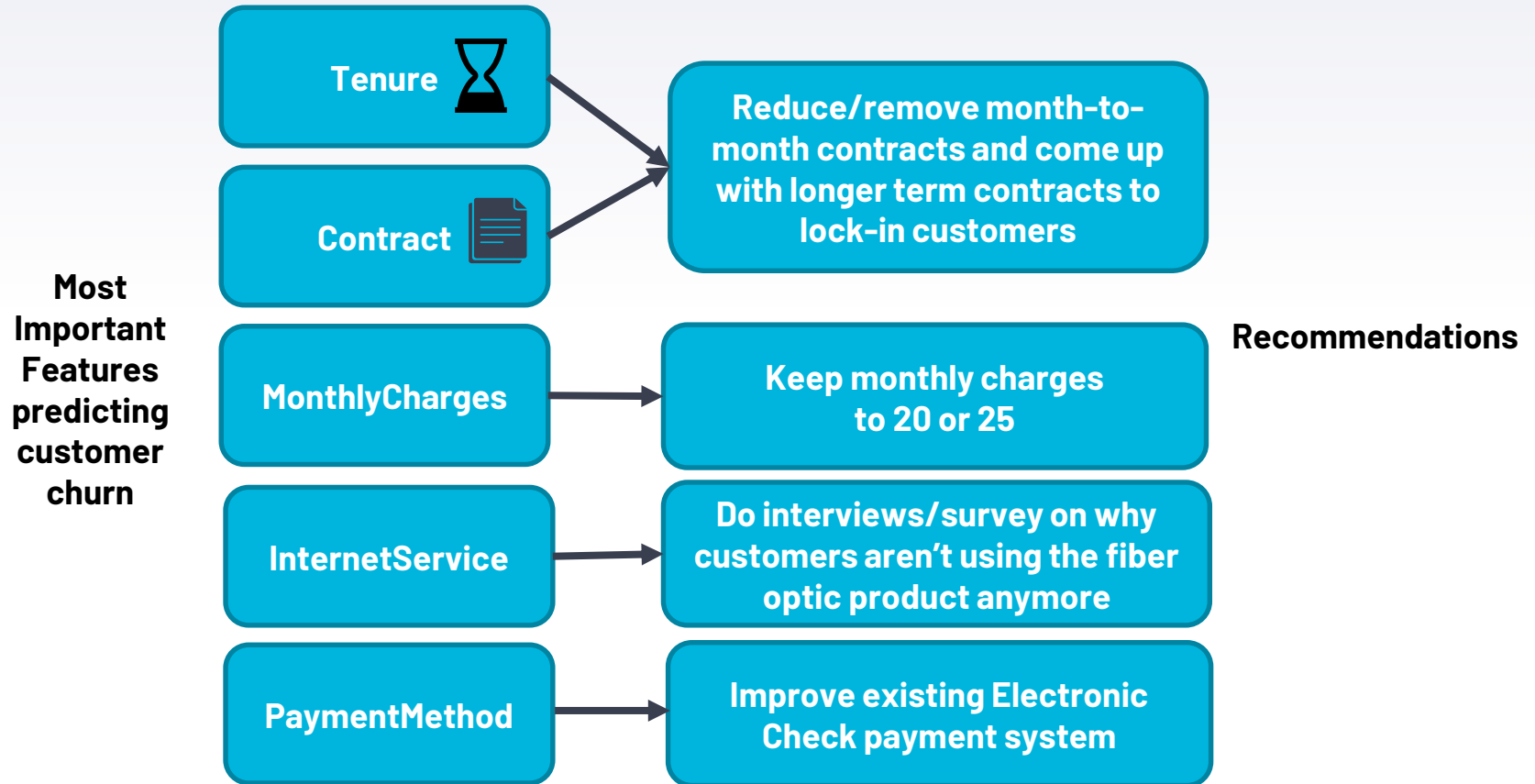
- 69% of customers who churned used fiber optic. This suggests that perhaps customers are not very happy using the company's existing fiber optic product and churn, switching to other alternatives.
- **Recommendation:** The company should do an investigation and get feedback on why customers aren't using their fiber optic product anymore, perhaps through interviews or a survey.

# PaymentMethod – Deep dive and Recommendation



- 57% of customers who churned used Electronic check to pay. This suggests that perhaps Electronic check may be a very cumbersome and inconvenient method for customers to pay, especially when the charges come in monthly.
- **Recommendation:** The company should look into their existing electronic check payment system and see if they can expedite the process.

# Summary of Recommendations





# THANK YOU!

