

Kaggle客戶流失率預測

Group 3

統計三 陳采宗
統計三 林盈盈
統計三 鄭雅云
統計三 沈冠宇

目錄

1 資料簡介

2 EDA

3 Model

4 Demo


5 結論


6 參考資料


1 資料簡介

Raw data

raw data 包含 7043 個觀測值以及 20 個變數


 Dataset




 1371




Telco Customer Churn

Focused customer retention programs

 BlastChar • updated 3 years ago (Version 1)

[Data](#) [Tasks](#) [Notebooks \(506\)](#) [Discussion \(10\)](#) [Activity](#) [Metadata](#)

[Download \(172 KB\)](#) [New Notebook](#) 

 Usability 8.8  License Data files © Original Authors  Tags business

Description

Context

"Predict behavior to retain customers. You can analyze all relevant customer data and develop focused customer retention programs." [IBM Sample Data Sets]

Input format

Input	Format
customerID	string
gender	binary
SeniorCitizen	binary
Partner	binary
Dependents	binary
tenure	integer
PhoneService	binary
MultipleLines	multinomial
InternetService	multinomial
OnlineSecurity	multinomial

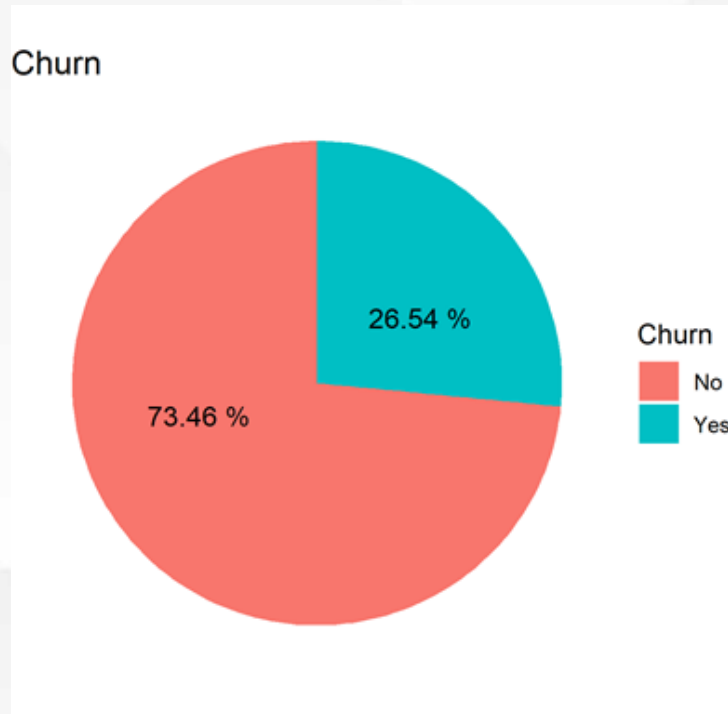
Input	Format
OnlineBackup	multinomial
DeviceProtection	multinomial
TechSupport	multinomial
StreamingTV	multinomial
StreamingMovies	multinomial
Contract	multinomial
PaperlessBilling	binary
PaymentMethod	multinomial
MonthlyCharges	numeric
TotalCharges	numeric

Data-preprocessing

- NA值處理：使用mice套件填補
- 原資料的SeniorCitizen欄位值為1、0，將其改為Yes、No

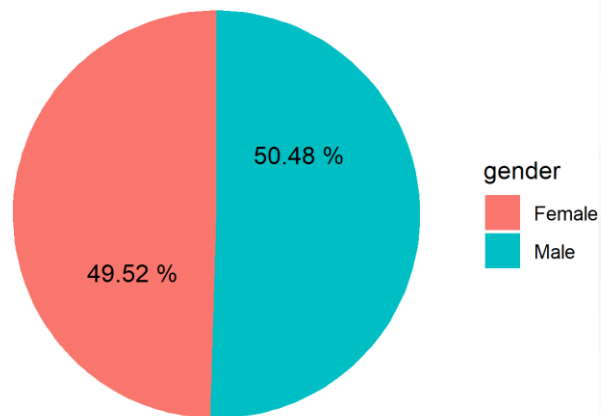
2 EDA

Churn

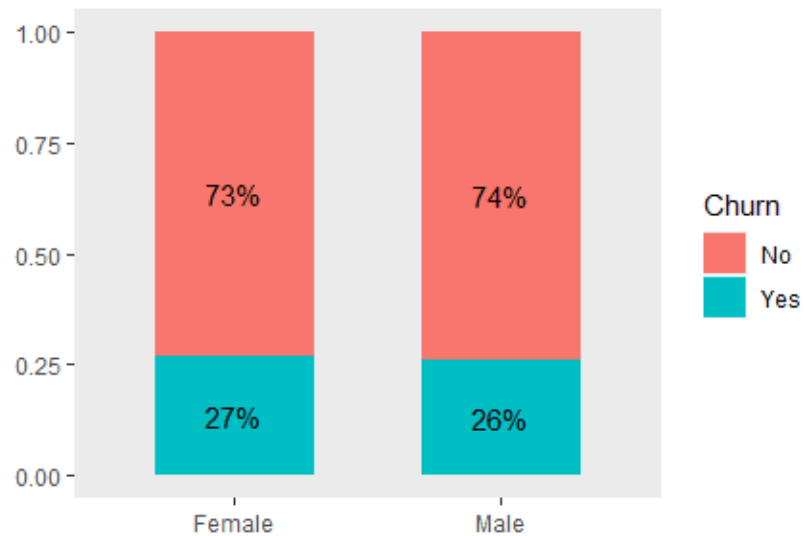


Gender & Churn-gender

Gender

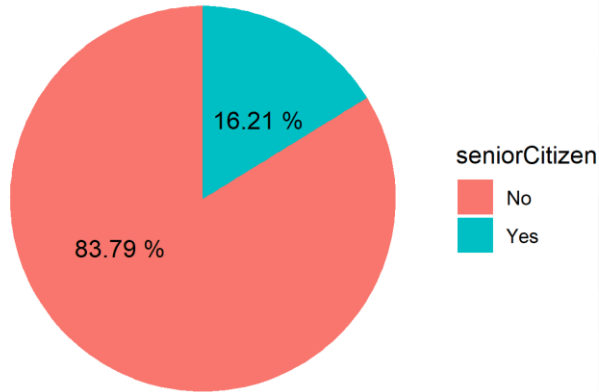


Gender

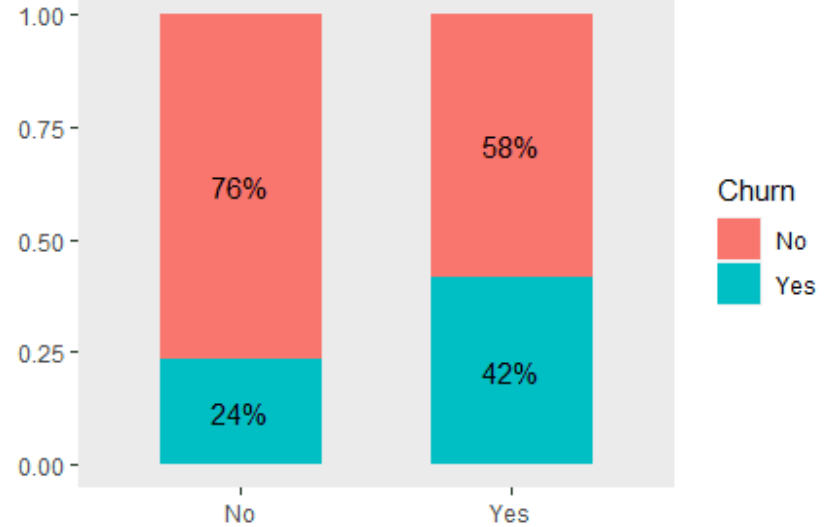


SeniorCitizen & Churn-SeniorCitizen

SeniorCitizen

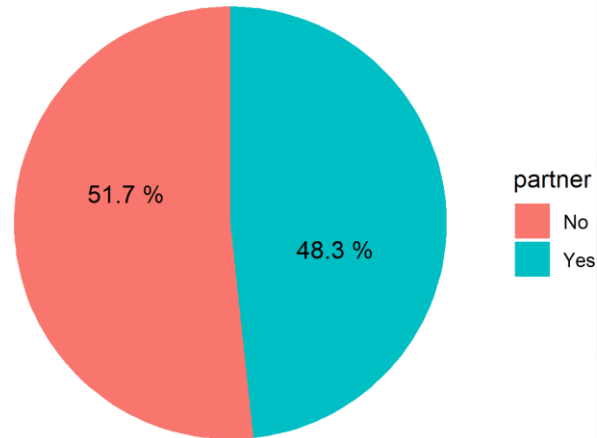


SeniorCitizen

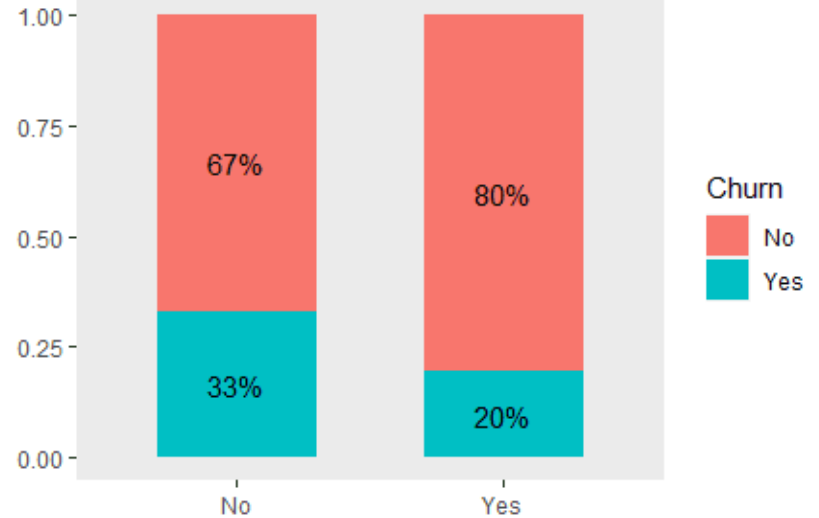


Partner & Churn-Partner

Partner

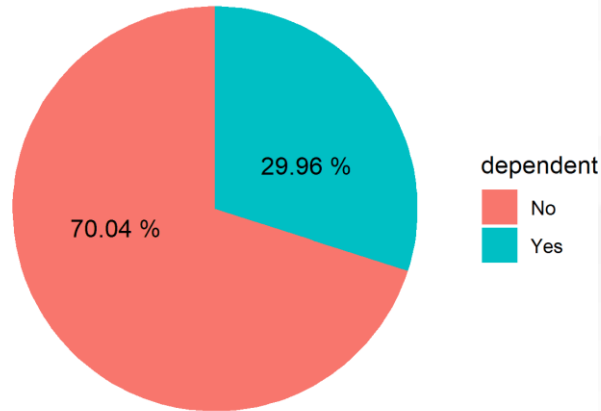


Partner

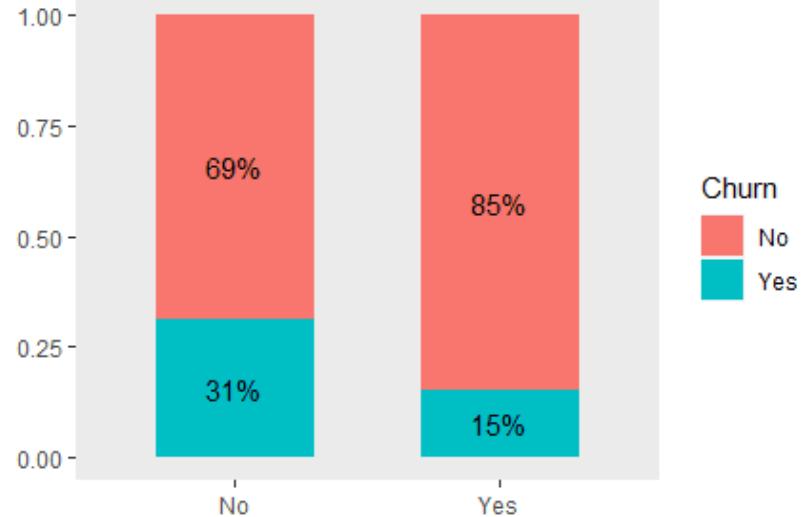


Dependents & Churn-Dependents

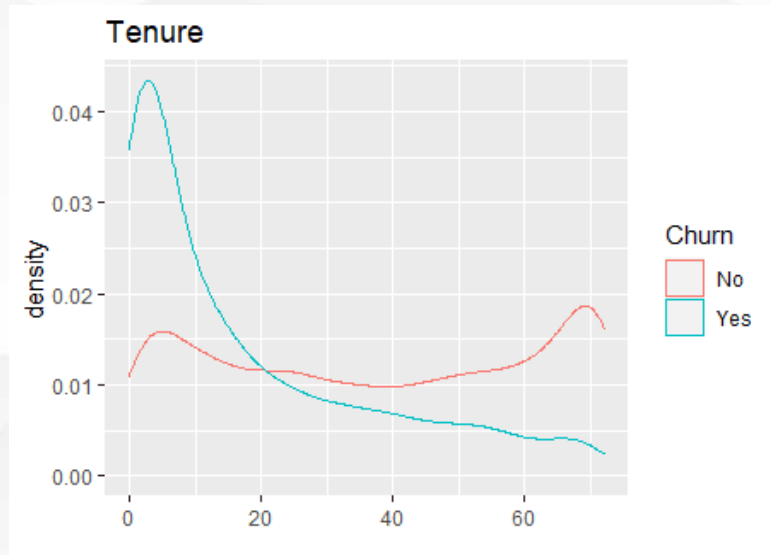
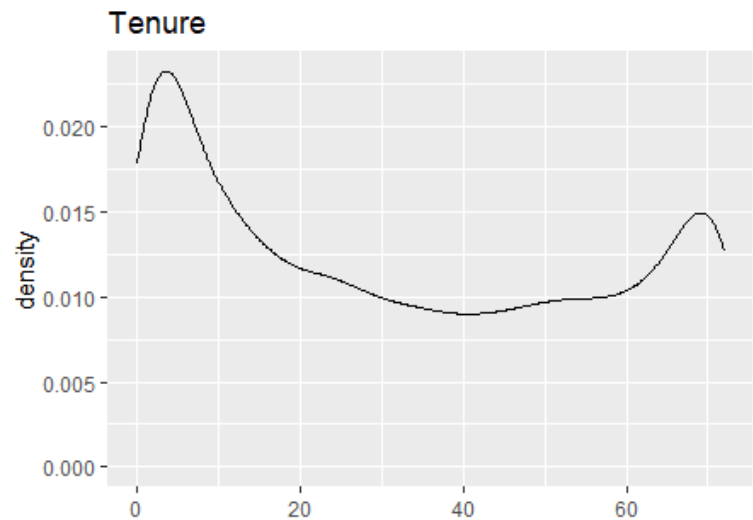
Dependent



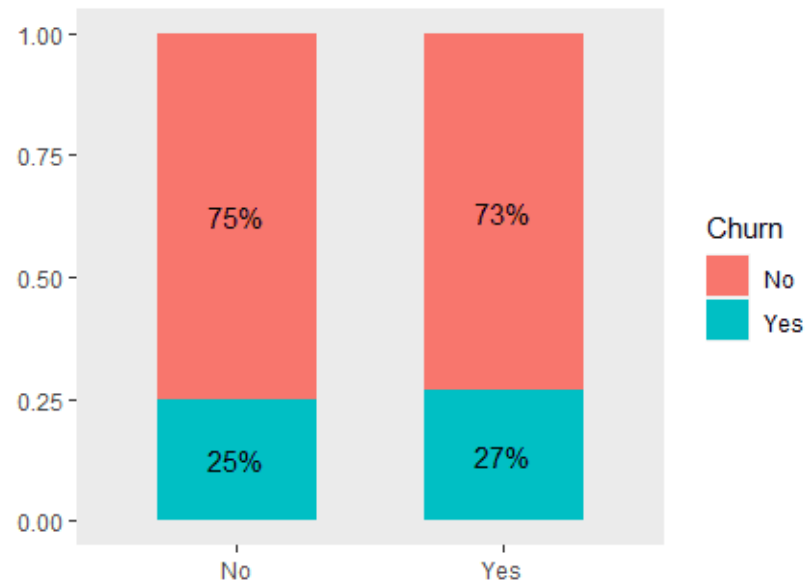
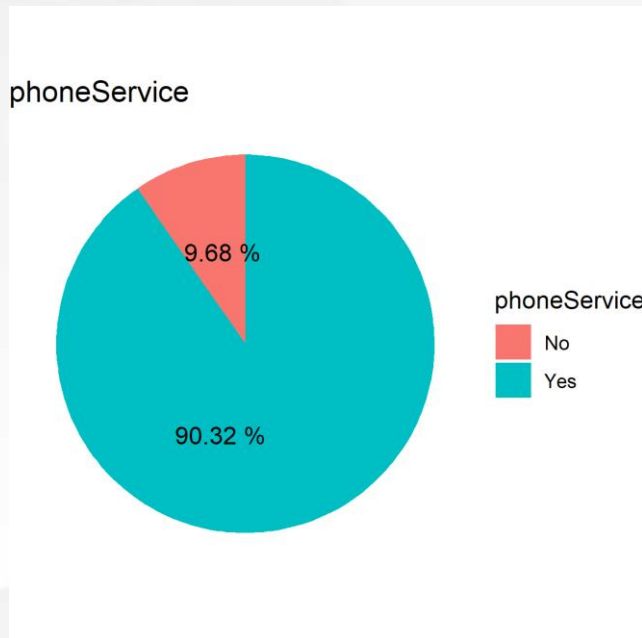
Dependents



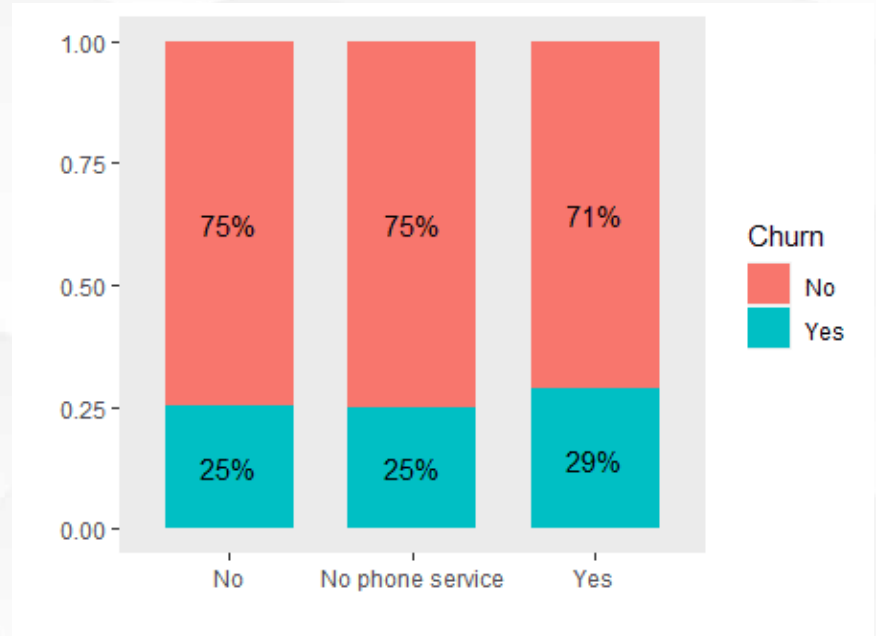
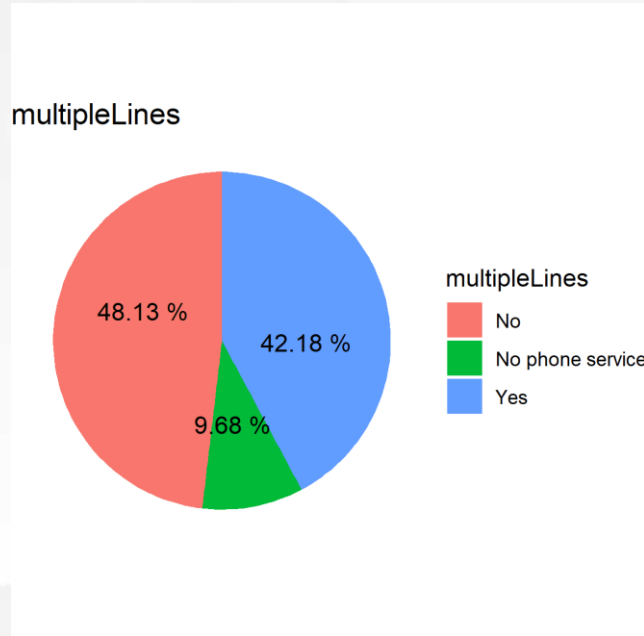
tenure & Churn-tenure



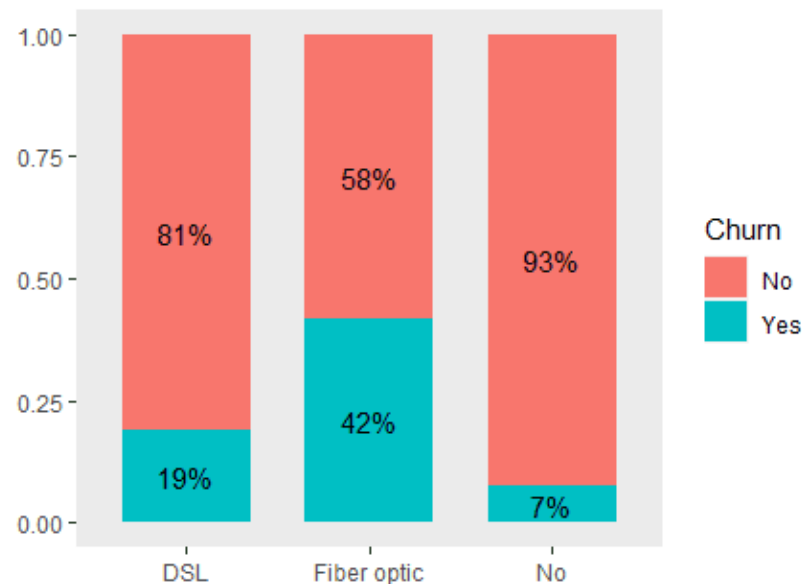
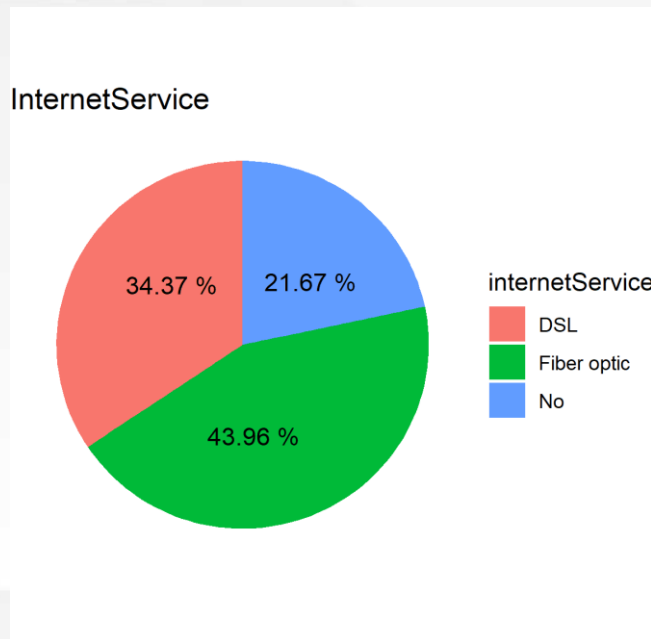
PhoneService & Churn-PhoneService



MultipleLines & Churn-MultipleLines

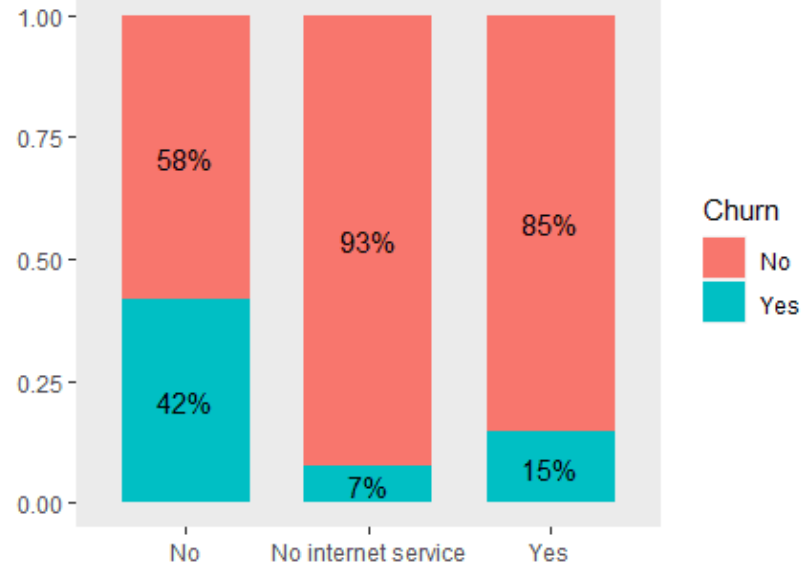
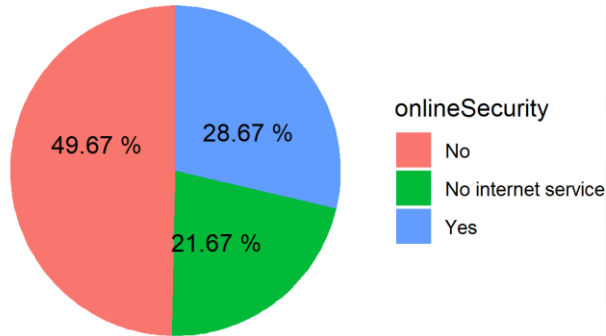


InternetService & Churn-InternetService

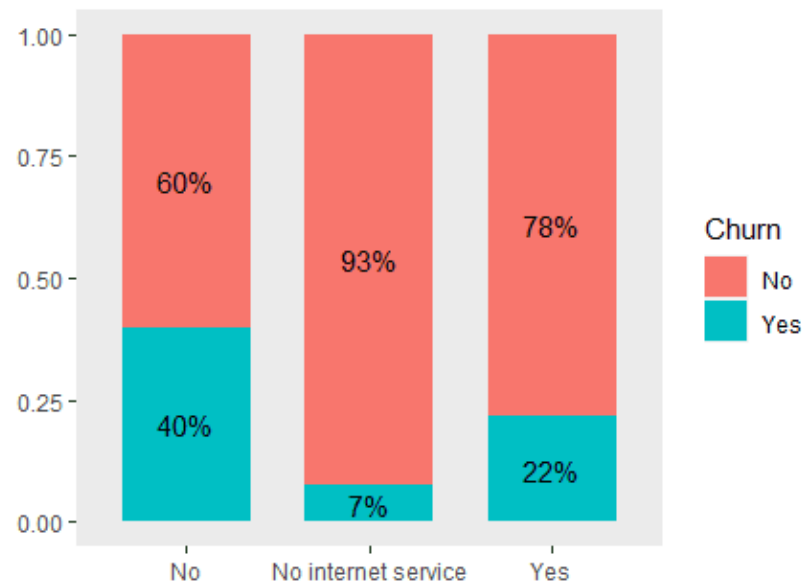
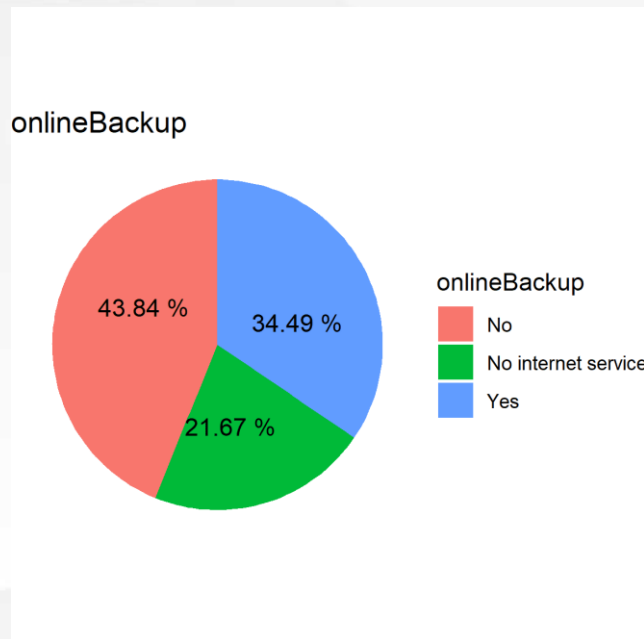


OnlineSecurity & Churn-OnlineSecurity

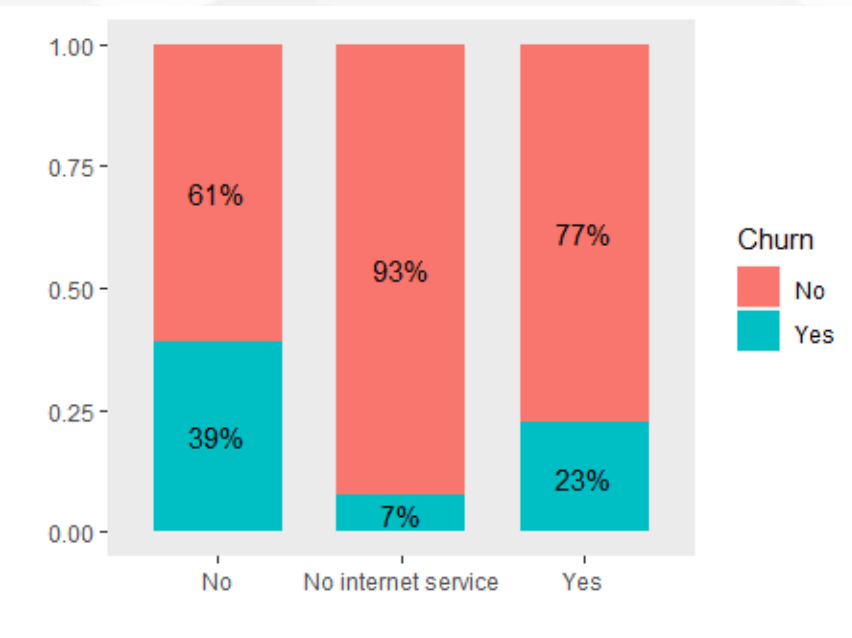
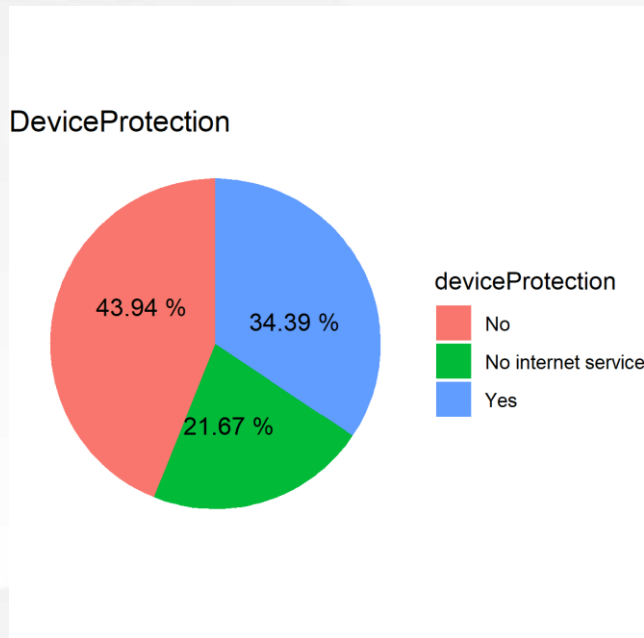
onlineSecurity



OnlineBackup & Churn-OnlineBackup

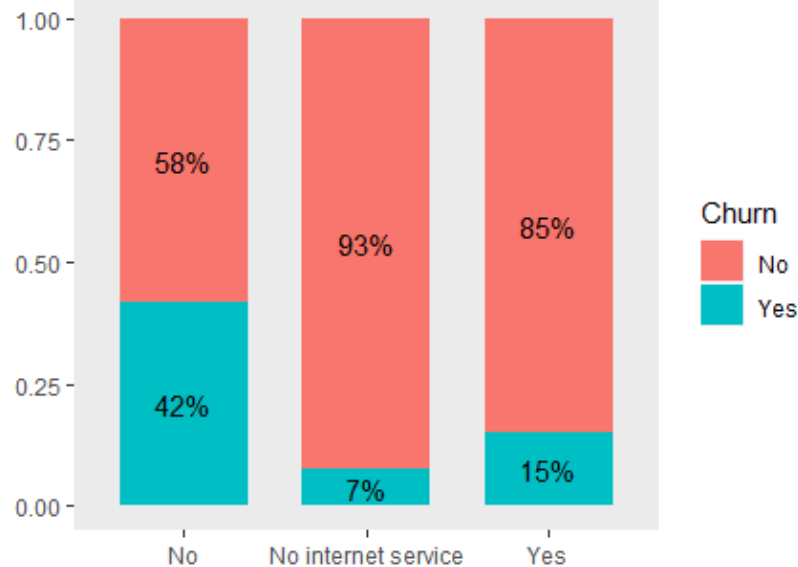
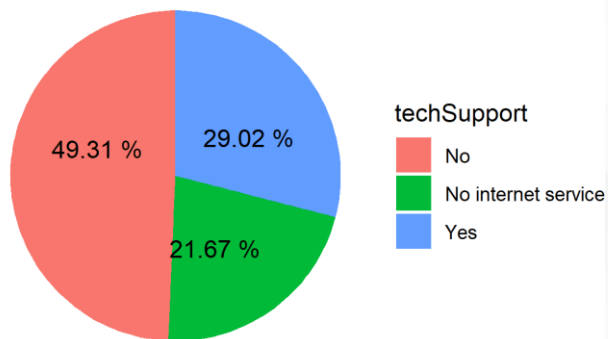


DeviceProtection & Churn-DeviceProtection



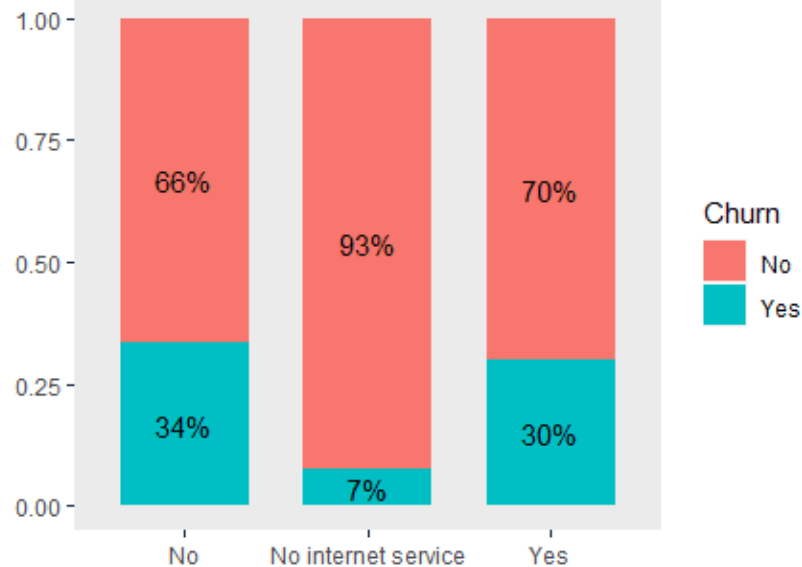
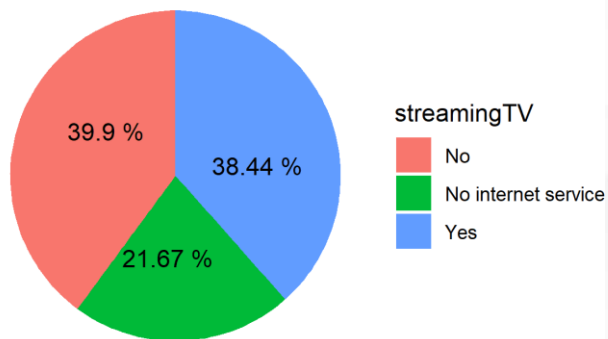
TechSupport & Churn-TechSupport

techSupport



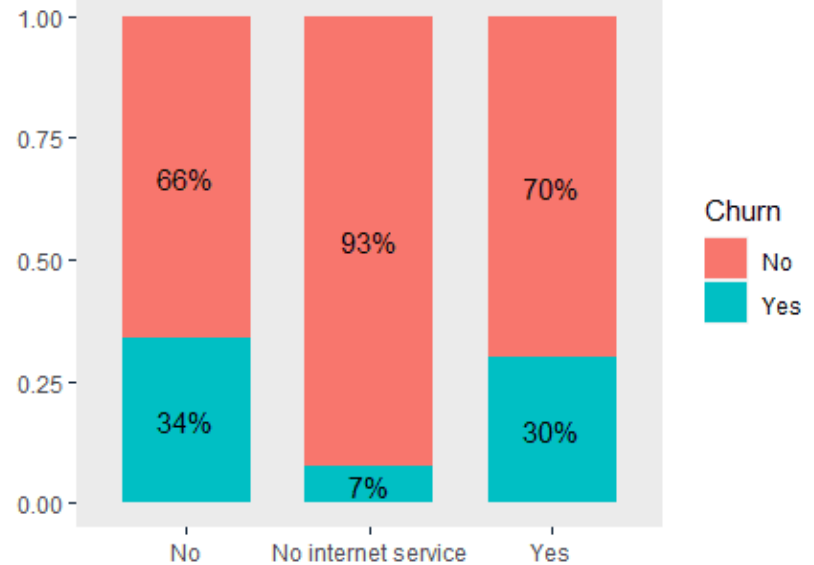
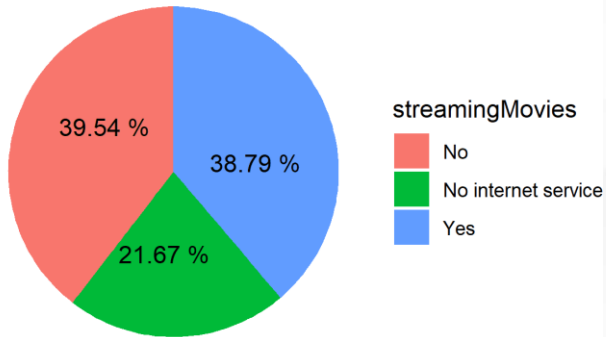
StreamingTV & Churn-StreamingTV

streamingTV



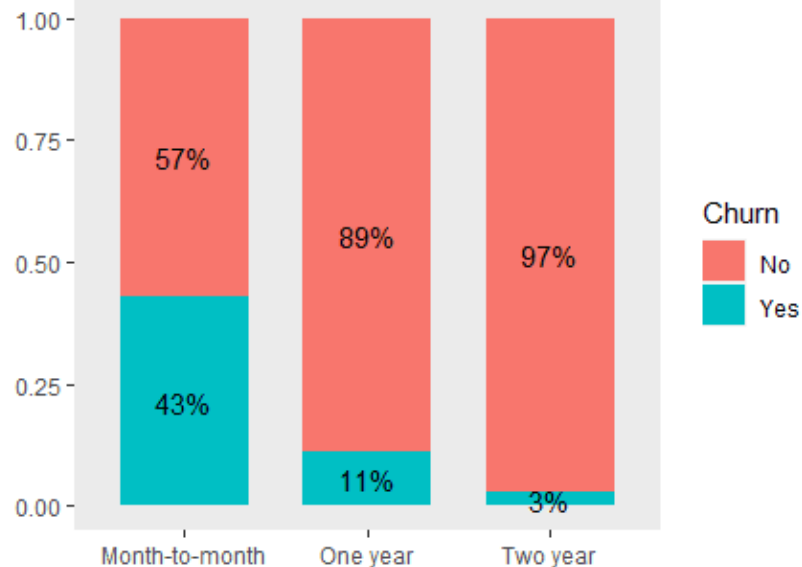
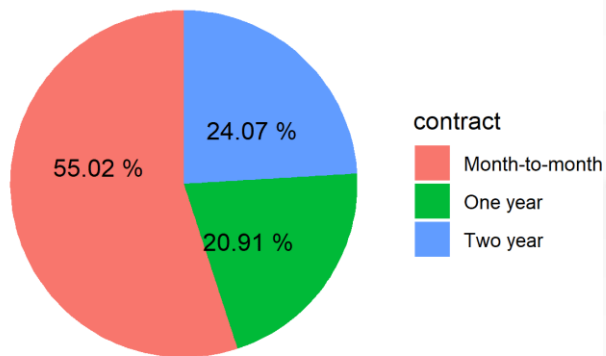
StreamingMovies & Churn-StreamingMovies

streamingMovies



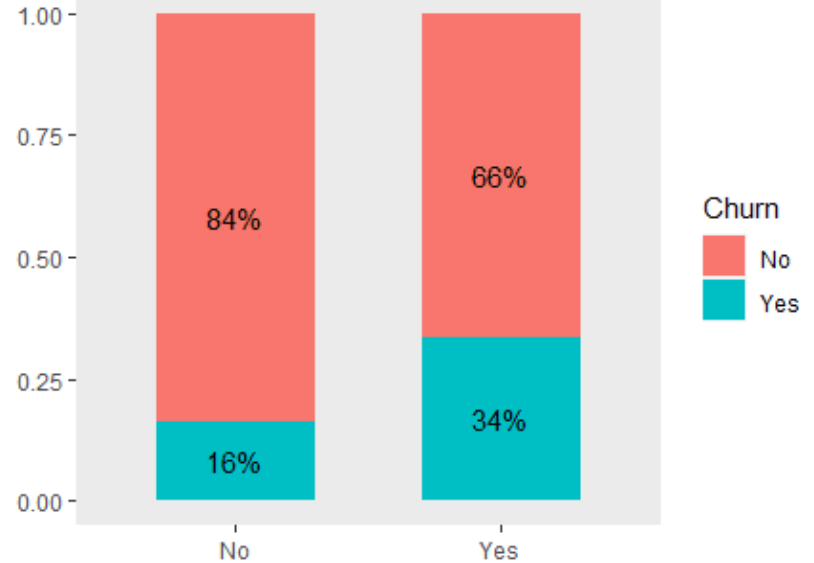
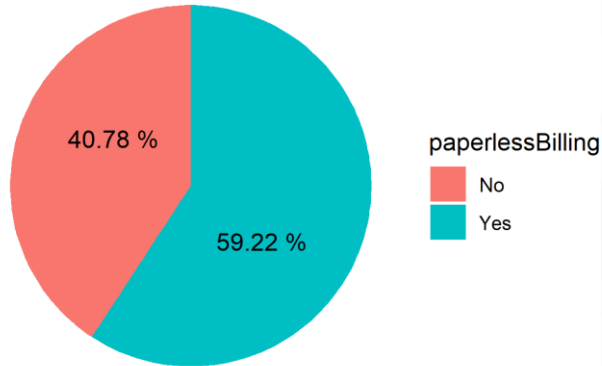
Contract & Churn-Contract

contract

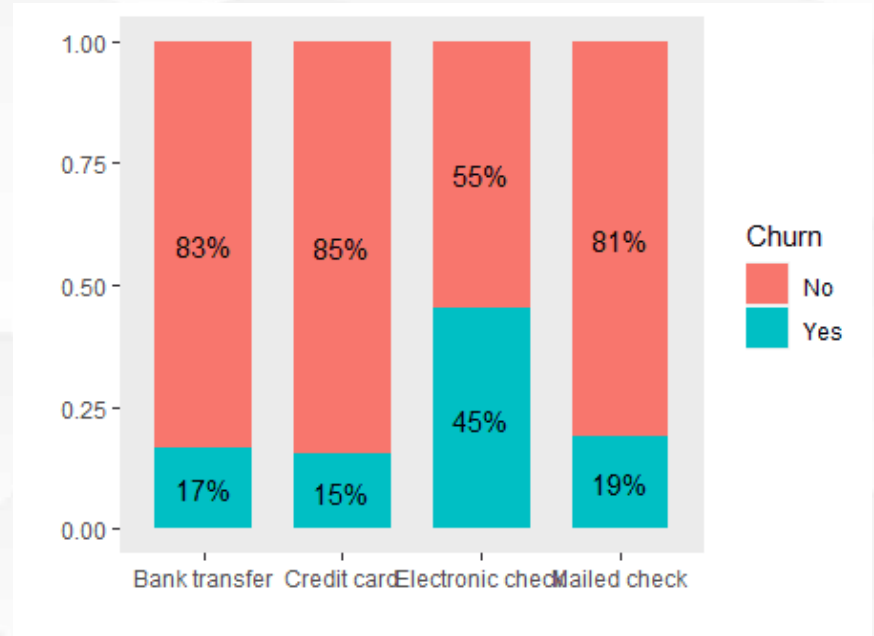
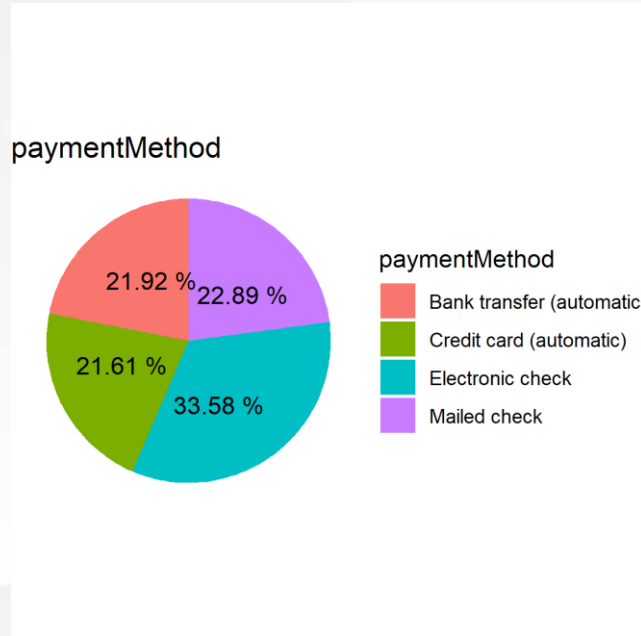


PaperlessBilling & Churn-PaperlessBilling

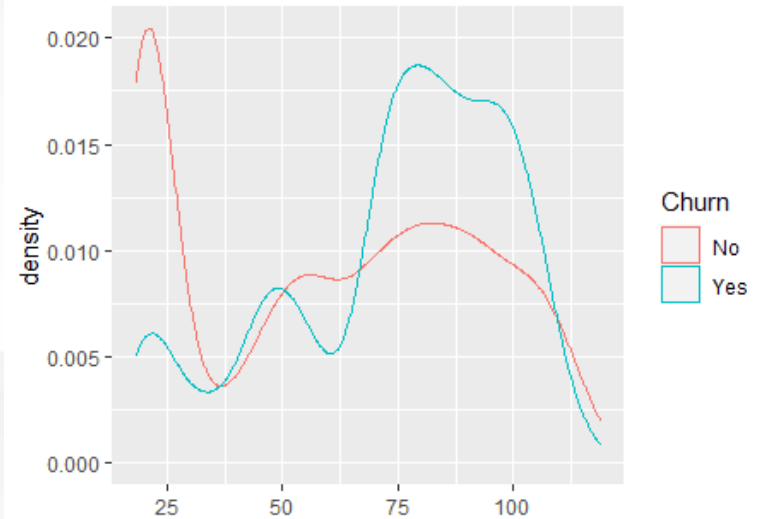
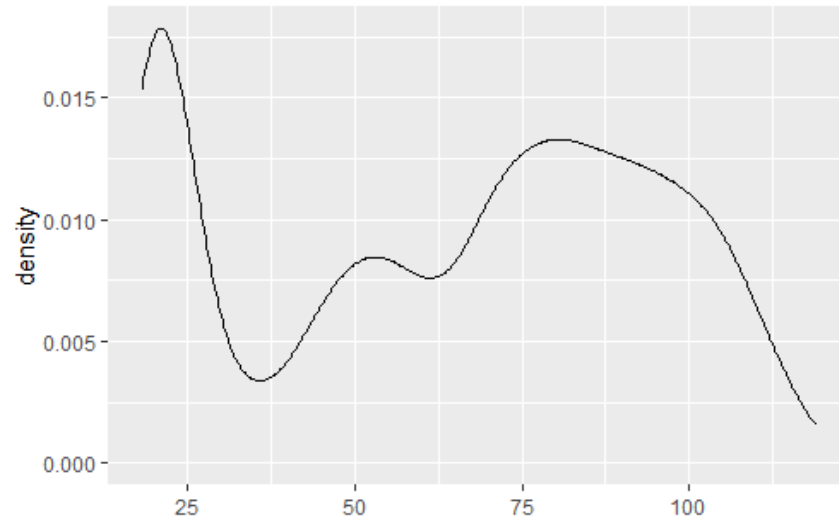
paperlessBilling



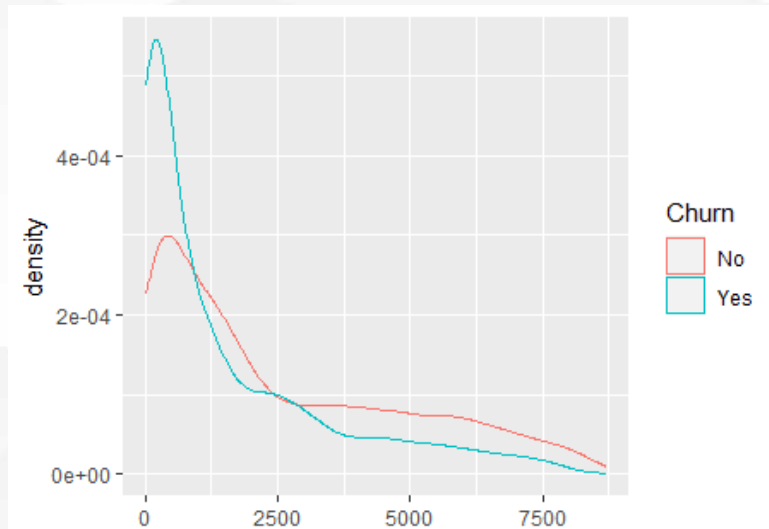
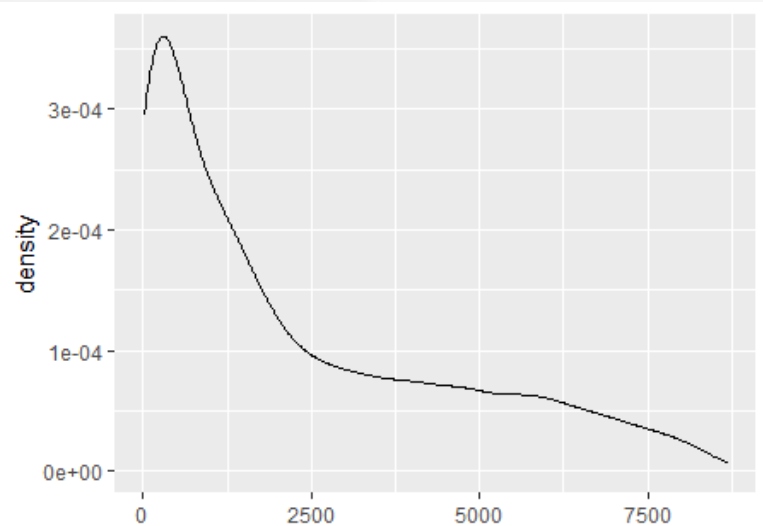
PaymentMethod & Churn-PaymentMethod



MonthlyCharges & Churn-MonthlyCharges



TotalCharges & Churn-TotalCharges



3 Model

Model 1 – Random Forest

➤ overfitting

Model 2 – Logistic Regression

- training/validation/training accuracy = 0.8 with all variables

Model 2 – Logistic Regression

- 挑選 logistic model 顯著之變數

Model 3 – Decision Tree

- 挑選 rpart 選用之變數



Stepwise

Data processing

- 連續型變數以pdf交叉點分割區間
- 新增變數：

Model

- 反覆挑選變數測試後test的accuracy仍無法突破0.8

4 Demo



Challenge



5 結論

結論

6 參考資料

參考資料

- <https://www.kaggle.com/blastchar/telco-customer-churn>
- <https://reurl.cc/3N1MgM>



Thank you!