Churn Analysis

Sergi Guimerà Roig

31 de desembre de 2022

Índex

1 Introducció

Les companyies de telecomunicacions no volen perdre clients. Per aquest motiu és habitual que contractin un especialista de Machine Learning per fer un anàlisi de dades i així poder predir quins són els clients que marxaran. Amb aquest anàlisi també s'extrauen les motius més habituals pels quals els clients decideixen abandonar la companyia i així poder fer una contraoferta i evitar perdre clients.
Una empresa dels sector ens ha contractat precisament per això i ens estipula 4 objectius principals:

- Quin tipus de persona marxa
- Que es pot fer per evitar-ho
- Quina gent no marxa
- Amb quina precisió podem predir la gent que marxa

Pels 3 primers punts farem user profiling amb els clients que marxen i els que no. Per l'últim objectiu utilitzarem mètodes supervisats per fer predicció de si el client marxa o roman a la companyia.

En aquest document es discutirà les decisions preses i el resultat d'aquestes.
Per aquesta finalitat s'ha utilitzat les dades de:

2 Metadata

Inicialment disposem d'una matriu de dades amb 7043 files i 38 columnes. És a dir 38 variables i 7043 clients. Aquestes 38 variables són:

Customer ID	Identificador del client. Està ordenat ascendentment.
Gender	Gènere del client (home/dona).
Age	Edat del client.
Married	Variable binària que ens indica si el client està casat.
Number of Dependents	Indica el nombre de persones dependents que conviuen amb el client (els dependents poden ser fills, pares, avis, etc.)
City	La ciutat on està la residència principal del client. Només ciutats de Califòrnia.
Zip Code	Zip Code de la residència principal del client.
Latitud	Latitud de la casa del client.
Longitud	Longitud de la casa del client.
Number of Referrals	Nombre de cops que el client ha fet referència de la companyia a amics o familiars.
Tenure in Months	Indica la quantitat total de mesos que el client ha estat a l'empresa al final del trimestre especificat anteriorment.
Offer	Variable categòrica que especifica l'última oferta de marketing que ha acceptat el client (None, Offer A, Offer B, Offer C, Offer D, Offer E).
Phone Service	Indica si el client té subscripció de línia telefònica a casa.
Avg Monthly Long Distance Charges	Indica els càrrecs mitjans de llarga distància del client, calculats fins al final del trimestre especificat anteriorment (si el client no està subscrit al servei de telefonia domiciliària, serà 0)
Multiple lines	Indica si el client està subscrit a múltiples línies telefòniques amb l'empresa: Sí, No (si el client no està subscrit al servei de telefonia domiciliària és No)
Internet Service	Indica si el client està subscrit al servei d'Internet amb l'empresa: Sí, No
Avg Monthly GB Download	Indica el volum mitjà de descàrrega del client en gigabytes, calculat fins al final del trimestre especificat anteriorment (si el client no està subscrit al servei d'Internet serà 0)
Online Security	Indica si el client està subscrit a un servei de seguretat en línia addicional que ofereix l'empresa: Sí, No (si el client no està subscrit al servei d'Internet, serà No)

Online Backup	Indica si el client està subscrit a un servei de còpia de seguretat en línia addicional proporcionat per l'empresa: Sí, No (si el client no està subscrit al servei d'Internet, serà No)
Device Protection Plan	Indica si el client està subscrit a un pla addicional de protecció de dispositius per als seus equips d'Internet proporcionat per l'empresa: Sí, No (si el client no està subscrit al servei d'Internet, serà No)
Premium Tech Support	Indica si el client està subscrit a un pla d'assistència tècnica addicional de l'empresa amb temps d'espera reduïts: Sí, No (si el client no està subscrit al servei d'Internet, serà No)
Streaming TV	Indica si el client utilitza el seu servei d'Internet per transmetre programes de televisió d'un tercer proveïdor sense cap cost addicional: Sí, No (si el client no està subscrit al servei d'Internet, serà No)
Streaming Movies	Indica si el client utilitza el seu servei d'Internet per reproduir pel·lícules d'un tercer proveïdor sense cap cost addicional: Sí, No (si el client no està subscrit al servei d'Internet, serà No)
Streaming Music	Indica si el client utilitza el seu servei d'Internet per reproduir música d'un tercer proveïdor sense cap cost addicional: Sí, No (si el client no està subscrit al servei d'Internet, serà No)
Unlimited Data	Indica si el client ha pagat una quota mensual addicional per tenir baixades/càrregues de dades il·limitades: Sí, No (si el client no està subscrit al servei d'Internet, serà No)
Contract	Indica el tipus de contracte actual del client: Mes a mes, Un any, Dos anys
Paperless Billing	Indica si el client ha optat per la facturació sense paper: Sí, No
Payment Method	Indica com el client paga la seva factura: Retirada bancària, Targeta de crèdit, xec enviat per correu
Monthly Charge	Indica el càrrec mensual total actual del client per tots els seus serveis de l'empresa
Total Charges	Indica els càrrecs totals del client, calculats fins al final del trimestre especificat anteriorment

Total Refunds	Indica els reemborsaments totals del client, calculats fins al final del trimestre especificat anteriorment
Total Extra Data Charges	Indica els càrrecs totals del client per baixades de dades addicionals per sobre de les especificades al seu pla, a finals del trimestre especificat anteriorment
Total Long Distance Charges	Indica els càrrecs totals del client per a llarga distància per sobre dels especificats al seu pla, al final del trimestre especificat anteriorment
Total Revenue	Indica els ingressos totals de l'empresa d'aquest client, calculats fins al final del trimestre especificat anteriorment (Càrrecs totals - Reemborsaments totals + Càrrecs totals de dades addicionals + Càrrecs totals de llarga distància)
Customer Status	Indica l'estat del client al final del trimestre: Churned, Stayed o Joined.
Churn Category	Una categoria d'alt nivell per a la raó del client per a la rotació, que es demana quan surt de l'empresa: Actitud, Competidor, Insatisfacció, Altres, Preu (relacionada directament amb la Churn Reason)
Churn Reason	El motiu específic d'un client per abandonar l'empresa, que es pregunta quan surt de l'empresa (directament relacionat amb Churn Category)

Sabem que les dades són de Califòrnia. Això també es pot veure si fem un scatter plot de Longitud en l'eix X i Latitud en l'eix Y ja que els punts agafen forma de Califòrnia.

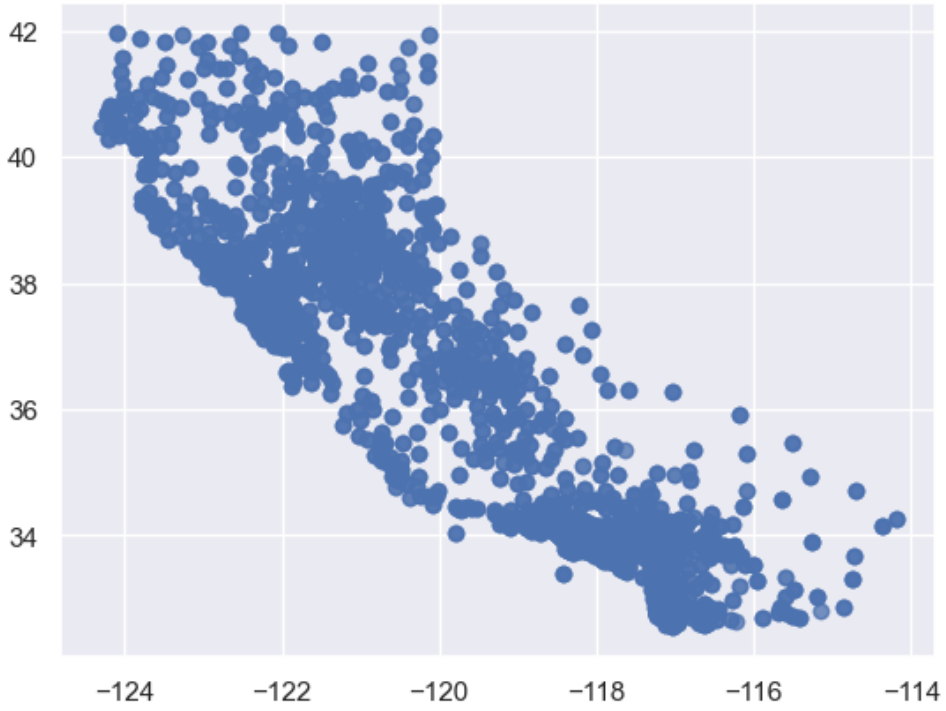

Figura 1: Plot Longitud-Latitud

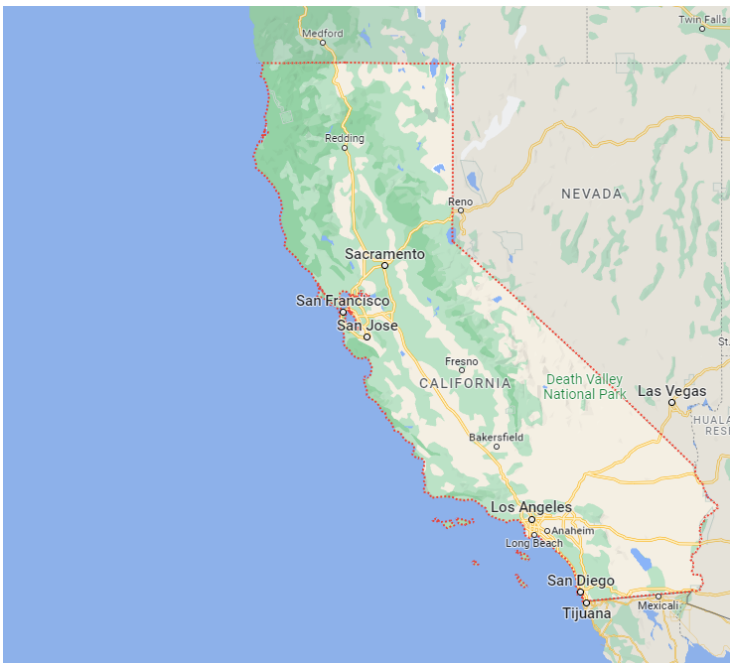

Figura 2: Califòrnia

3 Anàlisi i preprocessat de dades

Anteriorment hem vist que hi ha diferents variables que ens donen la mateixa informació, ja sigui de posició com City, Longitud-Latitud i Zip Code. O altres casos com Internet Type que ja ens abasta Internet Service. En el següent apartat veurem que fem amb aquests casos i eliminarem altres variables mentre visualitzem com es distribueixen aquestes. També tindrem en compte que les variables ens donin informació sobre la variable objectiu, si marxa o no.

3.1 Anàlisi correlacions

En aquest anàlisi estipularem que el llindar de decisió sobre si eliminar variables correlacionades és 0.8. És a dir que si tenim 2 variables que tenen una correlació amb valor absolut major o igual a 0.8, llavors haurem de eliminar o modificar les variables.

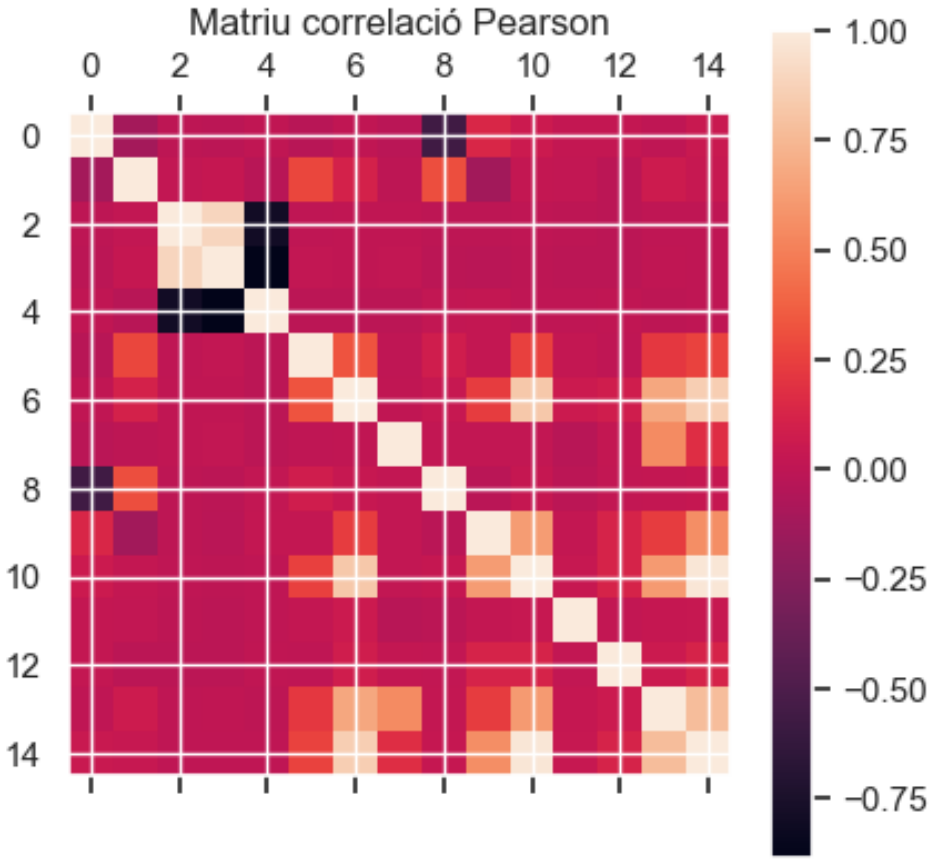

Figura 3: Matriu Correlació Pearson de les variables numèriques originals

Els identificadors de les variables per aquesta imatge són:
0: Age, 1: Number of Dependents, 2: Zip Code, 3: Latitude, 4: Longitude, 5: Number of Referrals, 6: Tenure in Months, 7: Avg Monthly Long Distance Charges, 8: Avg Monthly GB Download, 9: Monthly Charge, 10: Total Charges, 11: Total Refunds, 12: Total Extra Data Charges, 13: Total Long Distance Charges, 14: Total Revenue.
Podem veure que les variables bastant correlacionades són:

Longitude-latitude: El scatter plot és el que hem mostrat prèviament per comparar amb la forma de Califòrnia. Efectivament es nota que estan fortament correlacionades. El valor absolut supera el 0.8. En aquest cas, com tenim més variables que ens indiquen la zona eliminem Latitud i Longitud ja que al ser dues variables diferents ens pot portar problemes. Per exemple amb els models on utilitzem distàncies

com el KNN ja que al ser dues variables tindria més importància que si només utilitzem el Zip Code per exemple. No obstant això, Latitud i Longitud seria molt interessant d'utilitzar en els arbres de decisió i altres versions d'aquests ja que al fer un split amb una d'aquestes variables seria com agafar un mapa i partir-lo amb una línia. En un futur anàlisi es podria fer.
El Zip Code també està fortament correlacionat amb aquestes variables (0.895 per Latitud i -0.791 per Longitud), com eliminem tant Latitud com Longitud ja no ens hem de preocupar.
Aprofitem per dir que també eliminem la variable City ja que té moltes categories i ens hem decidit per confiar en Zip Code per donar informació posicional als models.

Age-Avg Monthly GB Download: Sembla bastant intuïtiu que en general, com més joves els clients més descarreguin d'internet. La correlació és -0.567 així que no modifiquem les variables de moment. Però mirem quina forma agafa scatter plot.

Figura 4: Scatter plot Age-Avg Monthly GB Download

Sembla que la forma està condicionada per com han entrat les dades a la base de dades ja que hi ha franges de descarrega i edat on no hi ha cap client. Això ho veurem millor quan analitzem la variable Avg Monthly GB Download i Age. Per altra part hi ha un quadrat de no observacions de gent de amb menys de 30 anys que es descarreguin menys de 25 GB, com si no existissin. Això podria ser perquè les dades són només un subset de les dades totals.

Total Revenue- Tenure in Months: Totes les variables que són totals estan fortament correlacionades amb Tenure in Months, ja que com més temps porten és més fàcil que s'hagin gastat més diners. Aquestes variables les modificarem; les dividirem per Tenure in Months perquè siguin atemporals. Les que ja tinguem versions atemporals com Monthly Charge amb Total Charges haurem de borrar-les.

Figura 5: Scatter plot Total Revenue - Tenure in Months

3.2 Anàlisi Univariat

Com només ens interessen les variables que la variable resposta sigui dependents d'elles, llavors podem utilitzar tests estadístics per mirar si les variables es distribueixen independentment. En el cas de les numèriques podem utilitzar ANOVA, T-test o Kruskal-Wallis. Aquest últim al ser no-paramètric, no s'ha de complir les assumpcions dels dos primers models (independència, homoscedasticitat, normalitat del errors i linealitat) i per tant podem utilitzar més lliurement. Per les categòriques podem utilitzar tests de chi quadrat i visualitzar taules de contingència.

Els missings de totes les variables no són dades mancants, sinó que estan mal codificades respecte la informació del meta-data. Els casos on necessites tenir un servei principal contractat (internet, línia de telefon) per tenir un servei secundari (Streaming TV/ Streaming music/ Online Security/...) o els casos on un client no abandona la companyia i llavors la churn reason/churn category, aquestes variables que depenen d'unes altres són NA quan no ho haurien de ser.
Llavors la estratègia és assignar el valor corresponent en cada cas.

Age

	Age
count	7043.00
mean	46.51
std	16.75
min	19.00
25%	32.00
50%	46.00
75%	60.00
max	80.00

Figura 6: Descripció Age

No hi ha cap missing value. La mitjana és més de 46 anys i és bastant més alta que la mitjana d'edat de nord-Amèrica (38.5), possiblement perquè no hi ha clients menors de 19 anys. Tampoc hi ha clients majors de 80 anys cosa que és sospitosa.

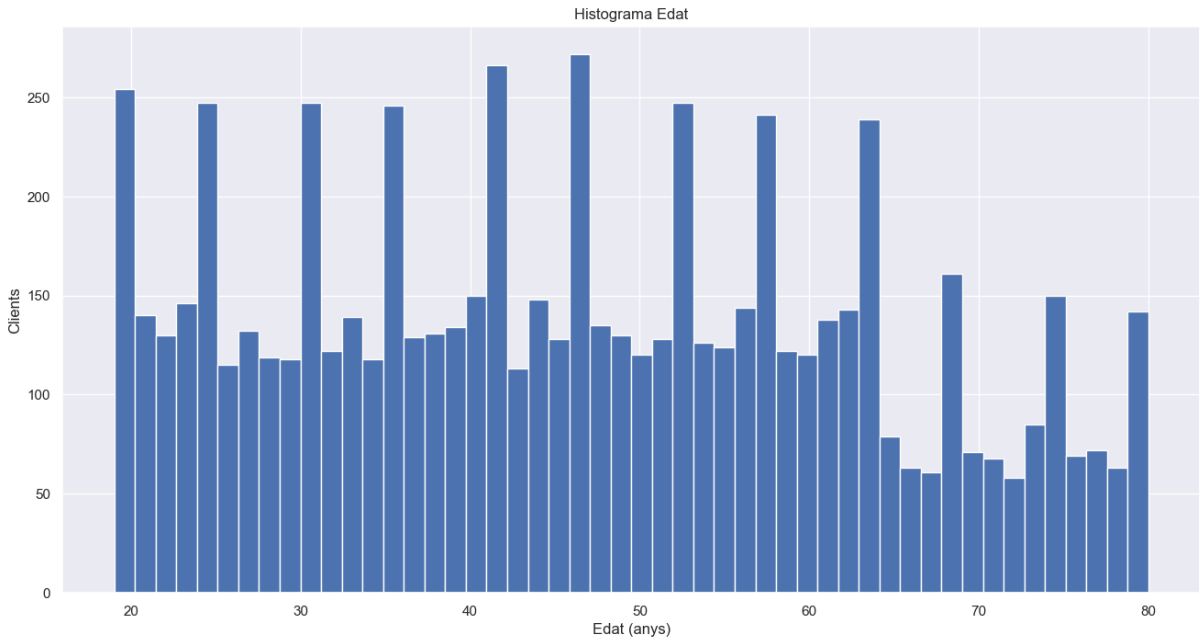

Figura 7: Histograma Age abans modificació

Podem veure que quan han entrat les dades han arrodonit cada 5 anys en bastants casos i per això ens queda aquesta forma dentada. Davant d'això tenim 2 estratègies. Deixar-ho així o aproximar-ho tot a múltiples de 5. La 1a estratègia té el problema que hi ha persones amb la edat real i persones amb l'edat aproximada. El problema de la 2a és que perdem precisió per la gent que si tenim la edat real. En aquesta ocasió aproximarem totes les edats. És a dir discretitzarem la edat a múltiples de 5.

	Age_disc
count	7043.000
mean	46.531
std	16.846
min	20.000
25%	30.000
50%	45.000
75%	60.000
max	80.000

Figura 8: Descripció Age modificada

Com només hem aproximat la descripció de la variable no ha canviat gaire. El canvi més notable és que ara el mínim és 20 anys.

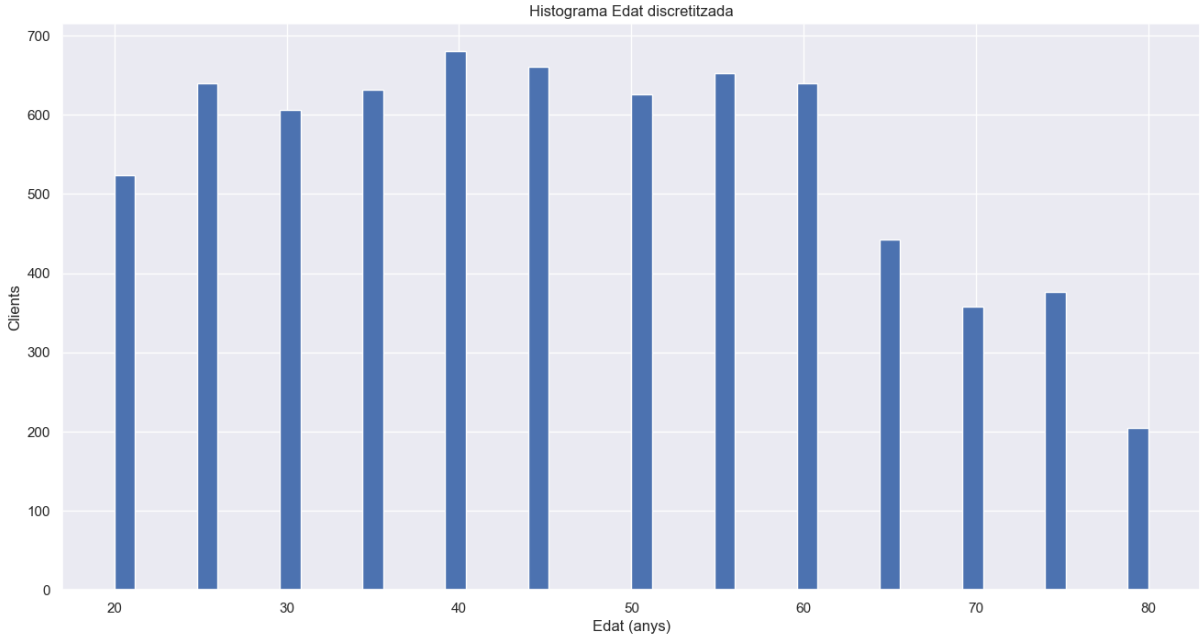

Figura 9: Histograma Age després modificació

La variable té forma gaussiana ni que és discreta.

Number of Dependents

	Number of Dependents
count	7043.000
mean	0.469
std	0.963
min	0.000
25%	0.000
50%	0.000
75%	0.000
max	9.000

Figura 10: Descripció Number of Dependents

Més de 3/4 de les observacions no tenen cap dependent i el màxim és 9. Potser aquesta variable es pot binaritzar si en té o no

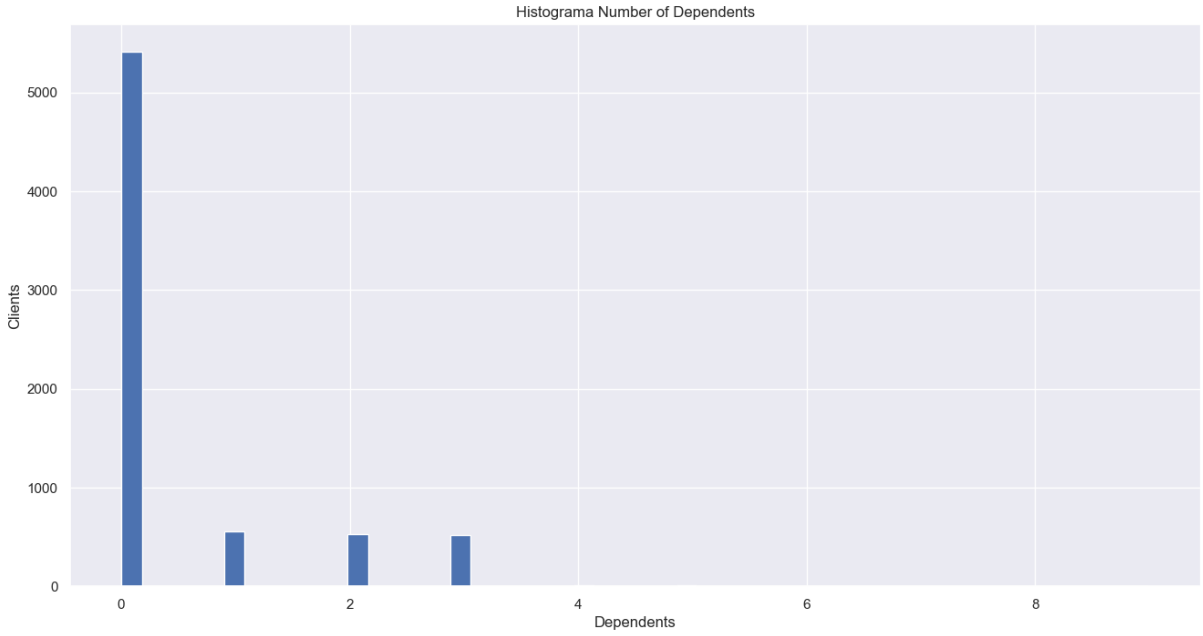

Figura 11: Histograma Number of Dependents abans modificació

Clients amb més de 3 n'hi ha tants pocs que ni es veu en el histograma. Decidim binaritzar la variable. 1 si té algun dependent i 0 d'altra forma. Obtenim la següent proporció.

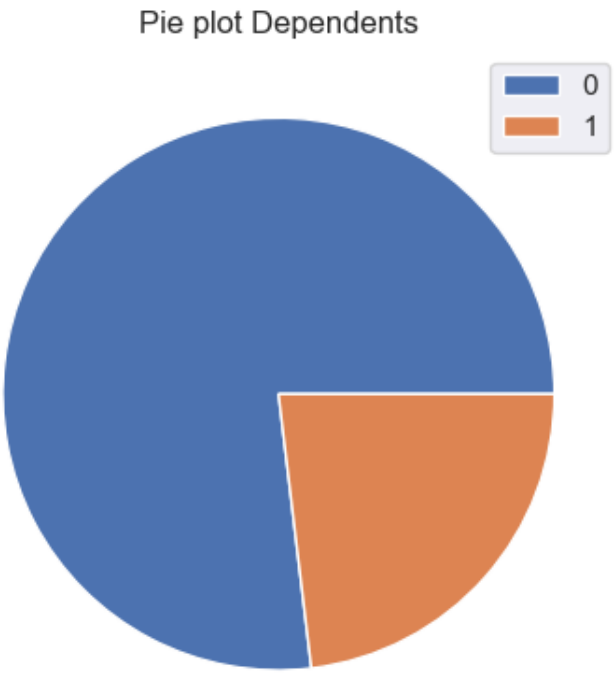

Figura 12: Pie plot Dependents

Zip Code

Zip Code és una variable que ens codifica els codis postals, la qüestió és que no és estrictament numèrica, això es pot veure ja que no té sentit sumar dos Zip Codes. No obstant llocs aprop tenen Zip Codes semblants, per això la deixem com a variable numèrica.

	Zip Code
count	7043.000
mean	93486.071
std	1856.768
min	90001.000
25%	92101.000
50%	93518.000
75%	95329.000
max	96150.000

Figura 13: Descripció Zip Code

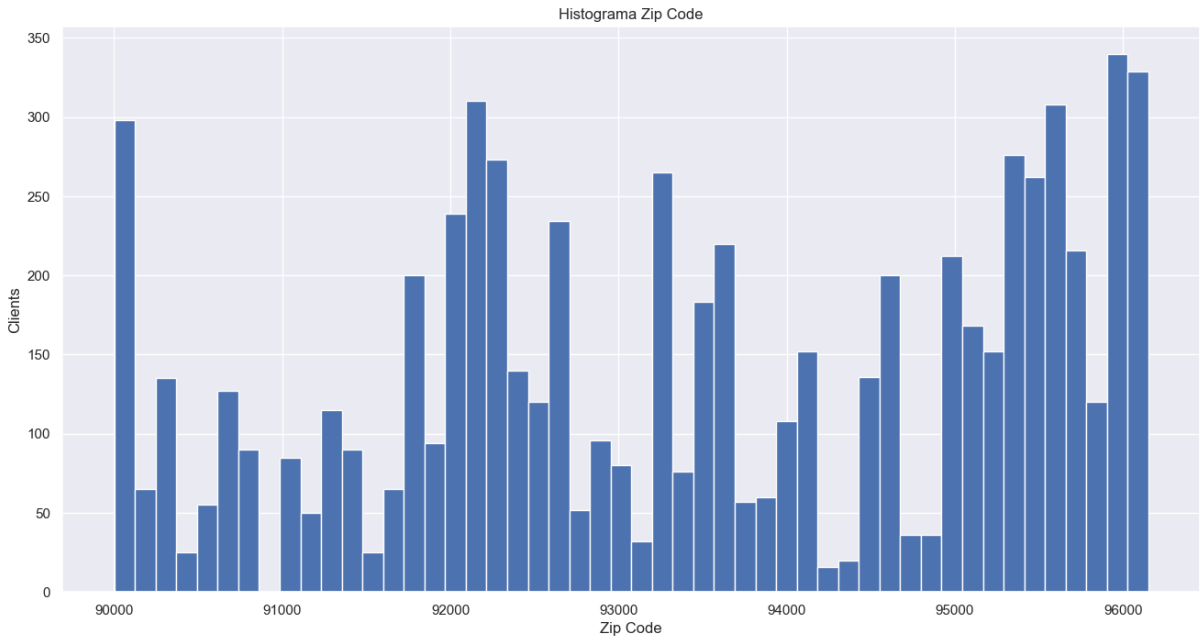

Figura 14: Histograma Zip Code abans modificació

Sabem que els primers números del Zip Code són més generals, i com més a la dreta més específic. Com hi ha molts Zip Codes diferents, agafem els 2 o 3 números de més a la dreta de cada Zip Code.

Figura 15: Histograma Zip Code 2 dígits

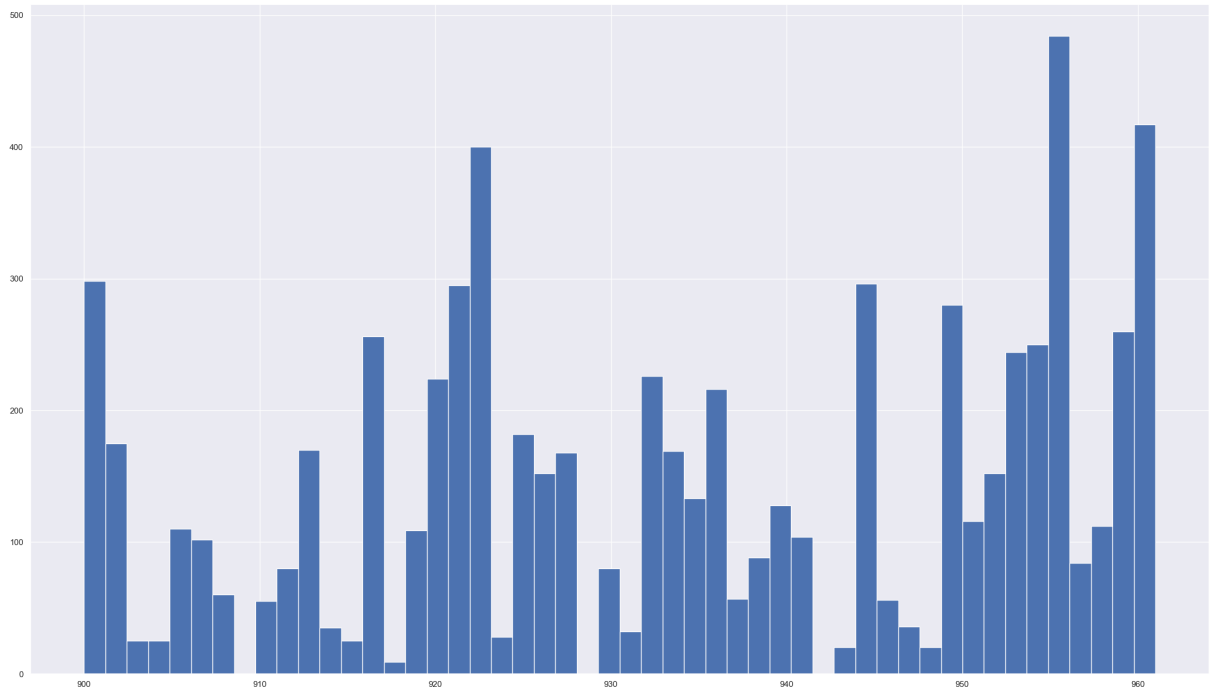

Figura 16: Histograma Zip Code 3 dígits

Si només agafem 2 digits de Zip Code ens queda 7 valors únics diferents, tenint en compte que el tractem com a numèric, 7 potser és molt poc així que agafem 3 digits.

	Zip Code 3
count	7043.000
mean	934.485
std	18.539
min	900.000
25%	921.000
50%	935.000
75%	953.000
max	961.000

Figura 17: Describe Zip Code 3 dígits

Number of Referrals

	Number of Referrals
count	7043.000
mean	1.952
std	3.001
min	0.000
25%	0.000
50%	0.000
75%	3.000
max	11.000

Figura 18: Descripció Number of Referrals

Més de la meitat de clients tenen 0 amics o familiars referits a la companyia. A més a més, els que han referit a més gent possiblement són clients antics.

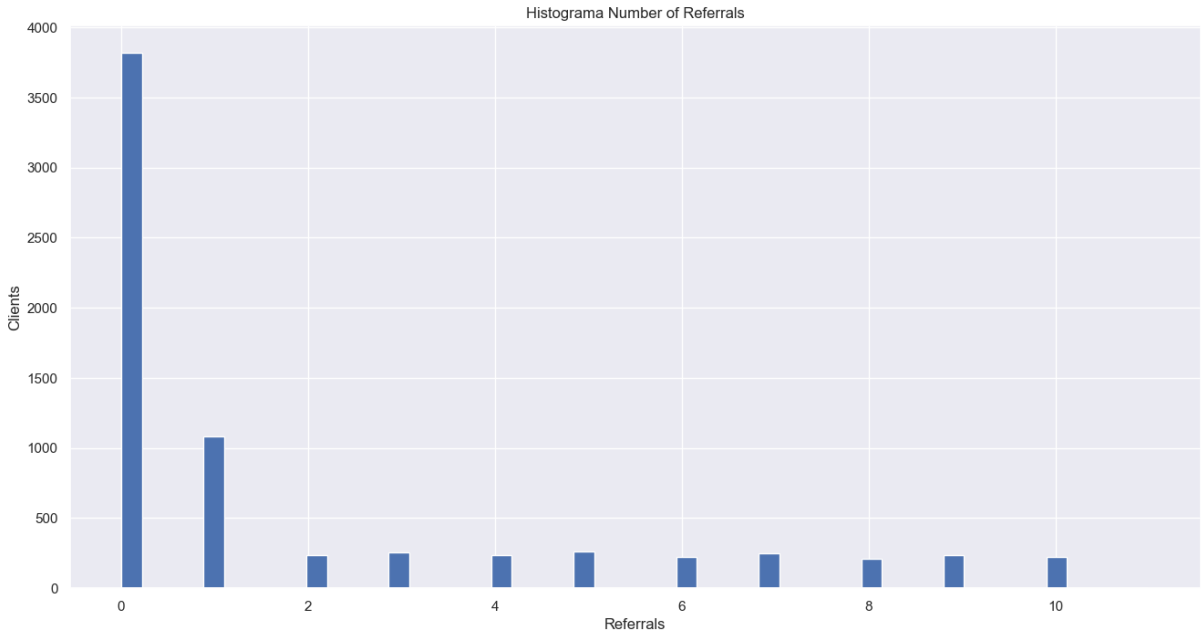

Figura 19: Histograma Number of Referrals abans modificació

Podem intentar balancejar la variable si la binaritzem, també eliminaríem una part de la relació amb el temps que té la variable. 1 si ha referit o 0 pel contrari. Decidim fer-ho i obtenim la següent proporció.

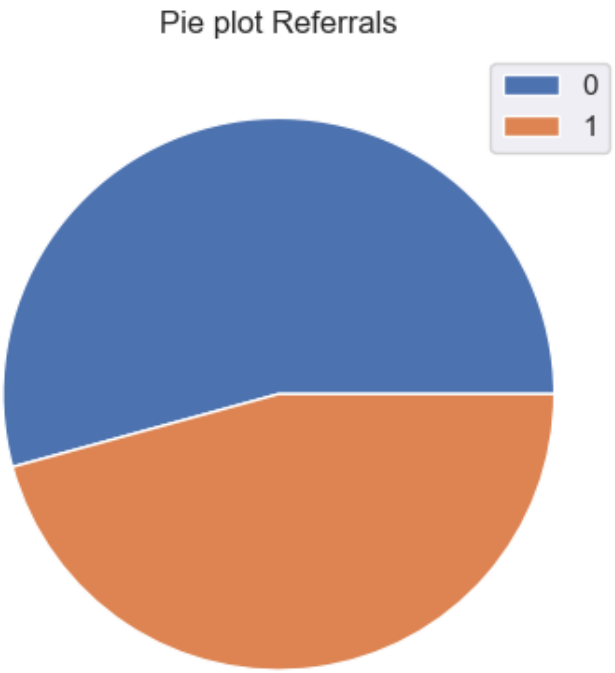

Figura 20: Pie plot Dependents

Tenure in Months

	Tenure in Months
count	7043.000
mean	32.387
std	24.542
min	1.000
25%	9.000
50%	29.000
75%	55.000
max	72.000

Figura 21: Descripció Tenure in Months

El màxim són 72 mesos que són 6 anys. Això podria ser degut a que abans la companyia no existia, no agafava data o només ens han cedit una part de les dades. El mínim és un mes, que són els clients nous.

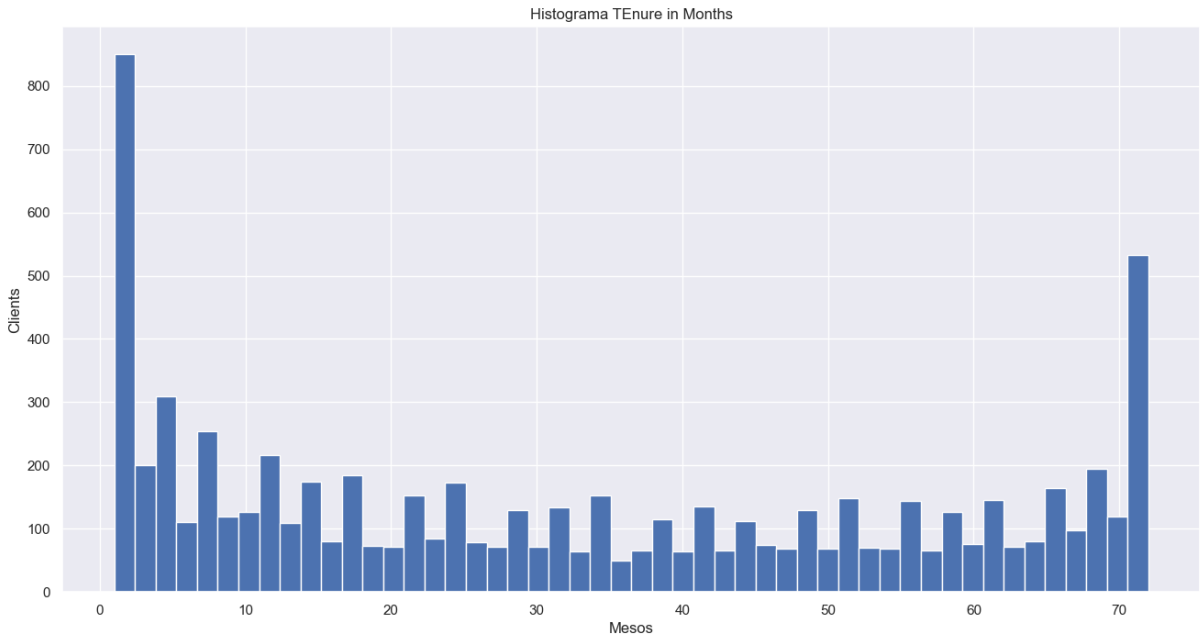

Figura 22: Histograma Tenure in Months

Podem veure que hi ha una acumulació de clients als dos extrems de l'histograma. En les extrems del rang. El mínim segurament són els clients nous que venen a provar. El màxim pot ser perquè han ajuntat tots els clients que portaven més temps. Apart podem veure que l'histograma està dentat, així que pot ser que les dades estiguin mig agrupades com en el cas de la edat. En aquest cop ho deixarem estar tal com està.

Avg Monthly Long Dist Charges

	Avg Monthly Long Distance Charges
count	6361.000
mean	25.421
std	14.200
min	1.010
25%	13.050
50%	25.690
75%	37.680
max	49.990

Figura 23: Descripció Avg Monthly Long Distance Charges

Hi ha missing values ja que el count és menor. El rang de la variable és [1, 50]. La mitjana i la mediana s'assemblen i estan al voltant de 25, el mig del rang.

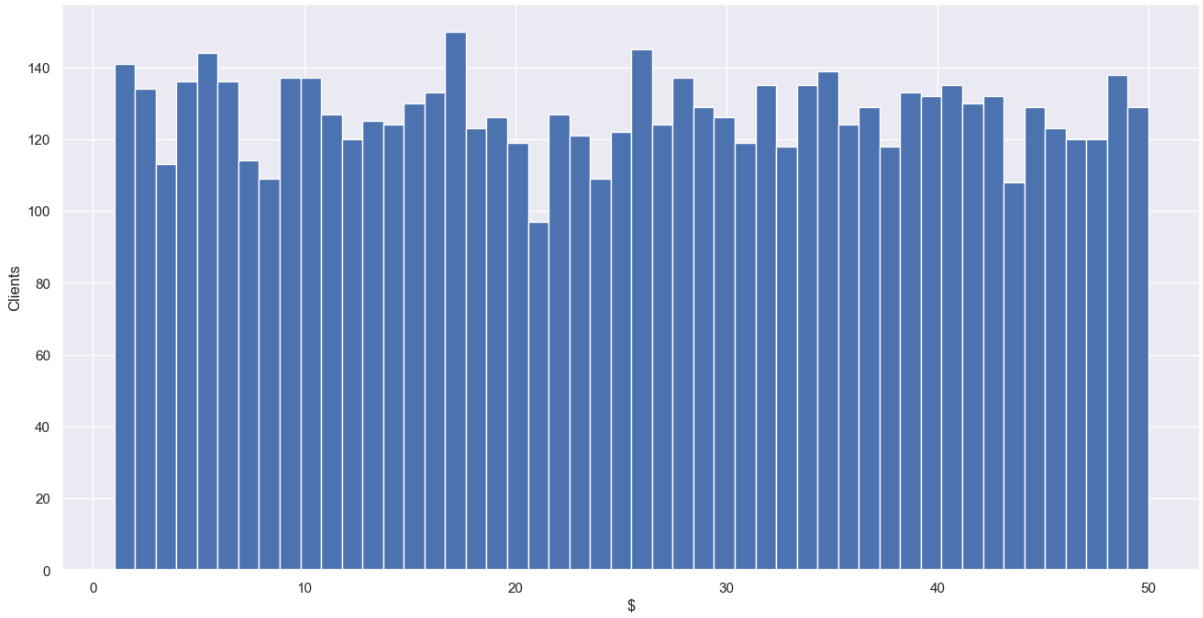

Figura 24: Histograma Avg Monthly Long Distance Charges

La distribució sembla uniforme.
Tots els clients que no tenen línia fixa en lloc de tenir un 0 són NA's. Posem els 0 que toca.

Avg Monthly GB Download

	Avg Monthly GB Download
count	5517.000
mean	26.190
std	19.587
min	2.000
25%	13.000
50%	21.000
75%	30.000
max	85.000

Figura 25: Descripció Avg Monthly GB Download

Hi ha missing values ja que el count és menor. El rang és de 2 GB a 85 GB amb una mitjana de 26.2.

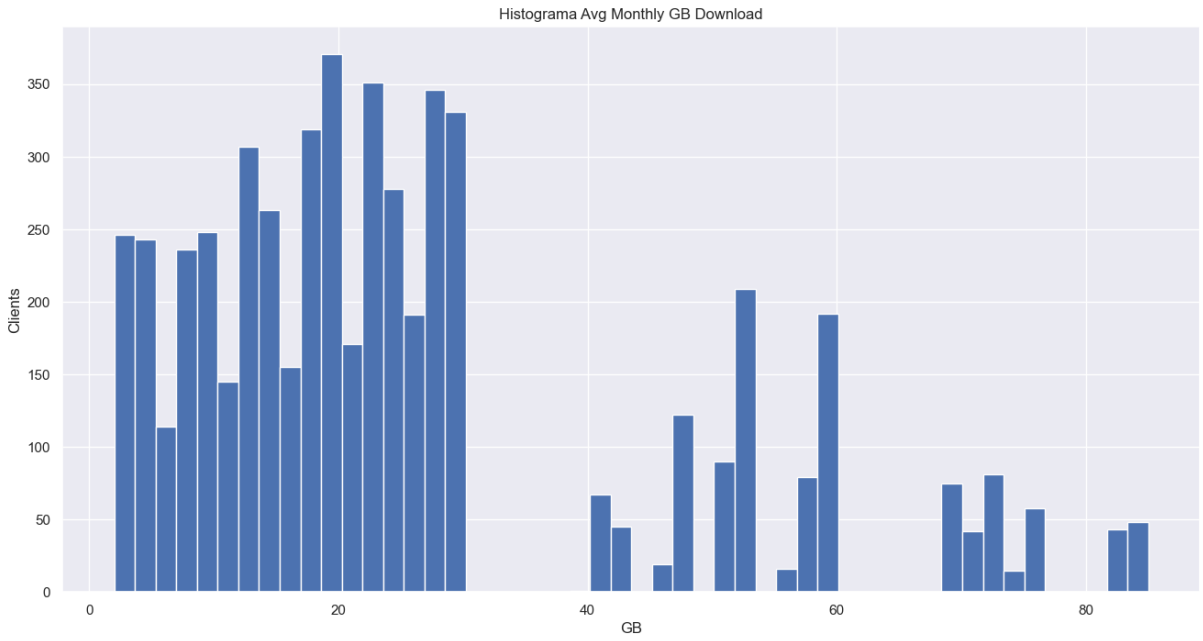

Figura 26: Histograma Avg Monthly GB Download

Aquí podem veure que l'histograma està dentat tal com hem predit en el primer anàlisi de correlacions. A més a més hi ha un forat de 30 a 40 GB, de 60 a 70 i tampoc hi ha 80s.

Els missings són tots aquells clients que no tenen internet contractat. Per imputar assignem 0 en aquests casos ja que no descarreguen res.

Monthly Charge

	Monthly Charge
count	7043.000
mean	63.596
std	31.205
min	-10.000
25%	30.400
50%	70.050
75%	89.750
max	118.750

Figura 27: Descripció Monthly Charge

No hi ha missings però hi ha valors per sota el 0. No té sentit perquè els valors haurien de ser estrictament positius. A més a més són valors entre -10 i -1 sense decimals. Segurament codifiquen alguna cosa. Nosaltres els haurem d'imputar o eliminar.

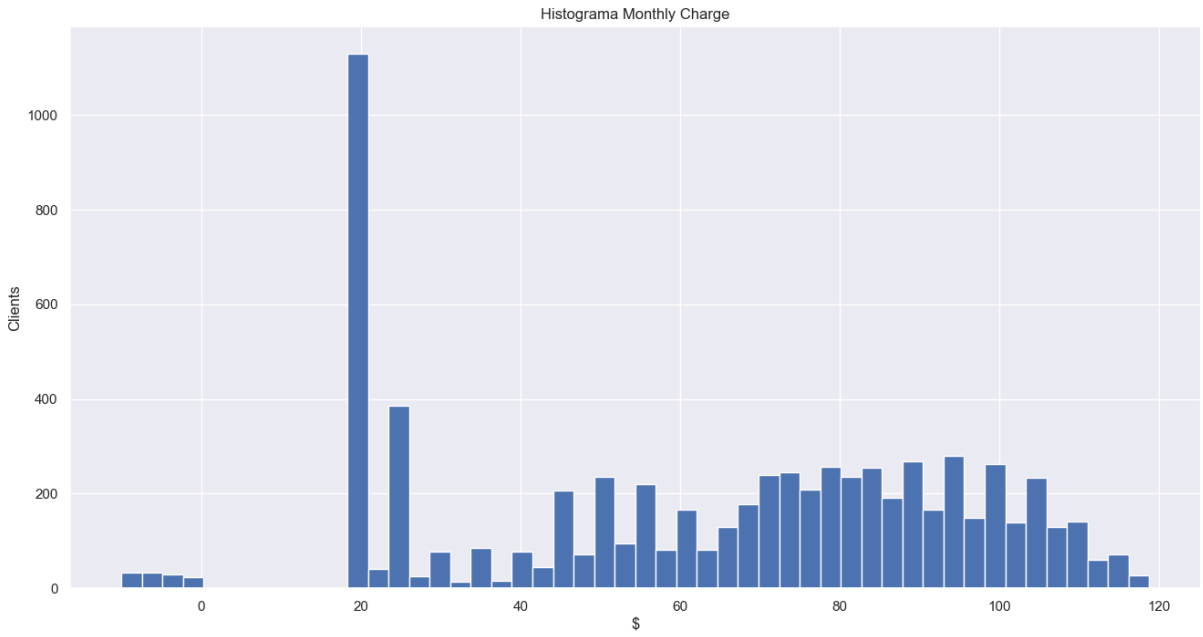

Figura 28: Histograma Monthly Charge

També veiem que els que paguen 20 són molts, això pot ser perquè és el preu bàsic sense res.

Com no sabem que vol dir la codificació, llavors assignem valors tal com si tinguéssim missing values. Una manera seria posar la mitjana de Monthly Charge, però els valors reals serien molt diferents sempre que caiguessin en un dels extrems del rang de valors de la variable. Podríem utilitzar un model de predicció per imputar, però com millor resultats doni el model més difícil haurà estat entrenar-lo. En aquest cas, com disposem de Total Charges i Tenure in Months. Podem calcular l'average monthly charge i assignar-lo en les dades anòmales. Si mirem la diferència entre aquests valors i Montlhy Charge pels clients que no hem d'imputar podem fer el següent histograma.

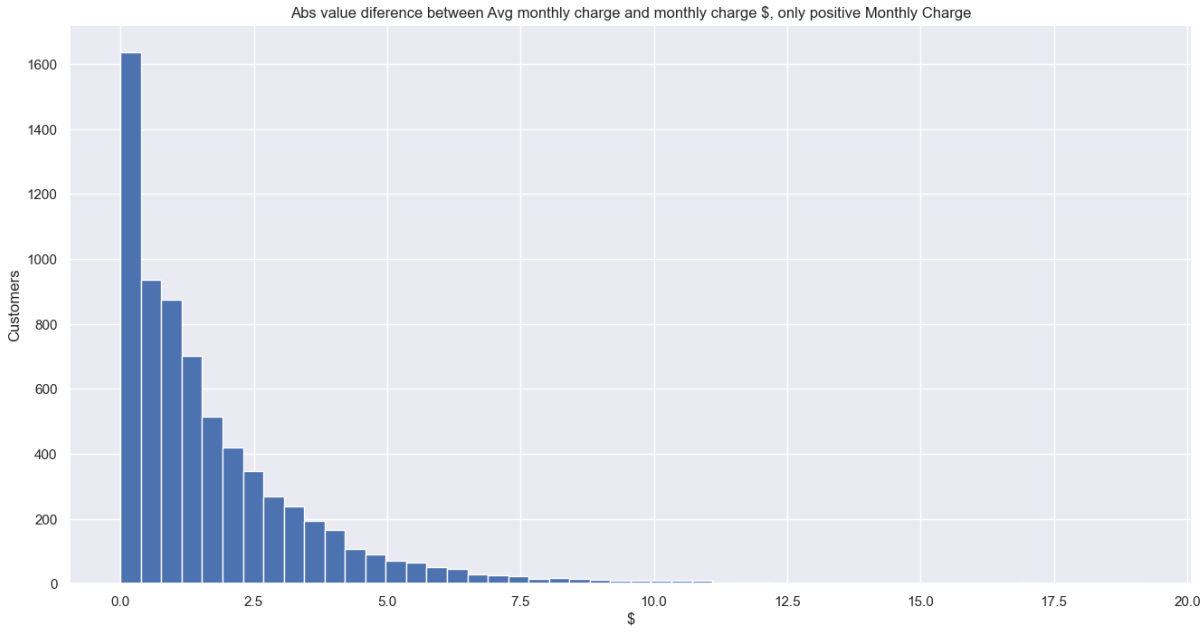

Figura 29: Histograma error imputació

Veiem que en la majoria de clients falla poc i que sempre està per sota de 20. Sembla un bon mètode així que l'utilitzem.

	Monthly Charge
count	7043.000
mean	64.761
std	30.095
min	16.750
25%	35.500
50%	70.300
75%	89.850
max	120.336

Figura 30: Descripció Monthly Charge imputat

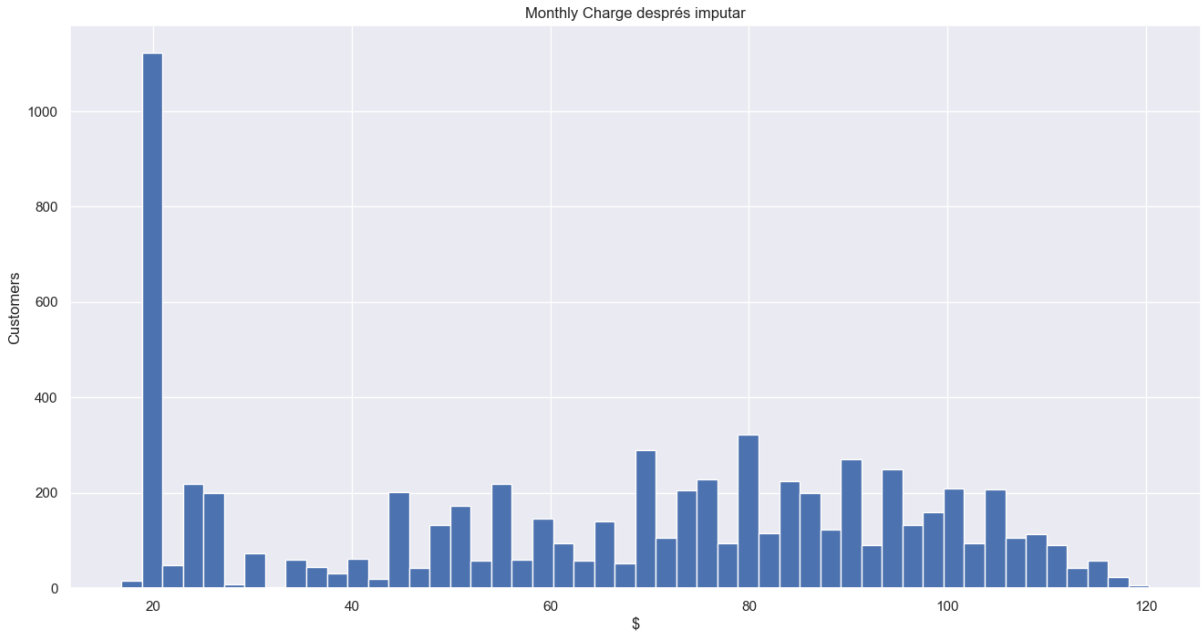

Figura 31: Histograma Monthly Charge imputat

Ja no hi ha valors negatius, ara el mínim és 16.75 i el màxim 120.

Total Charge

	Total Charges
count	7043.000
mean	2280.381
std	2266.220
min	18.800
25%	400.150
50%	1394.550
75%	3786.600
max	8684.800

Figura 32: Descripció Total Charge

El rang és de 18.8 a 8684.8.

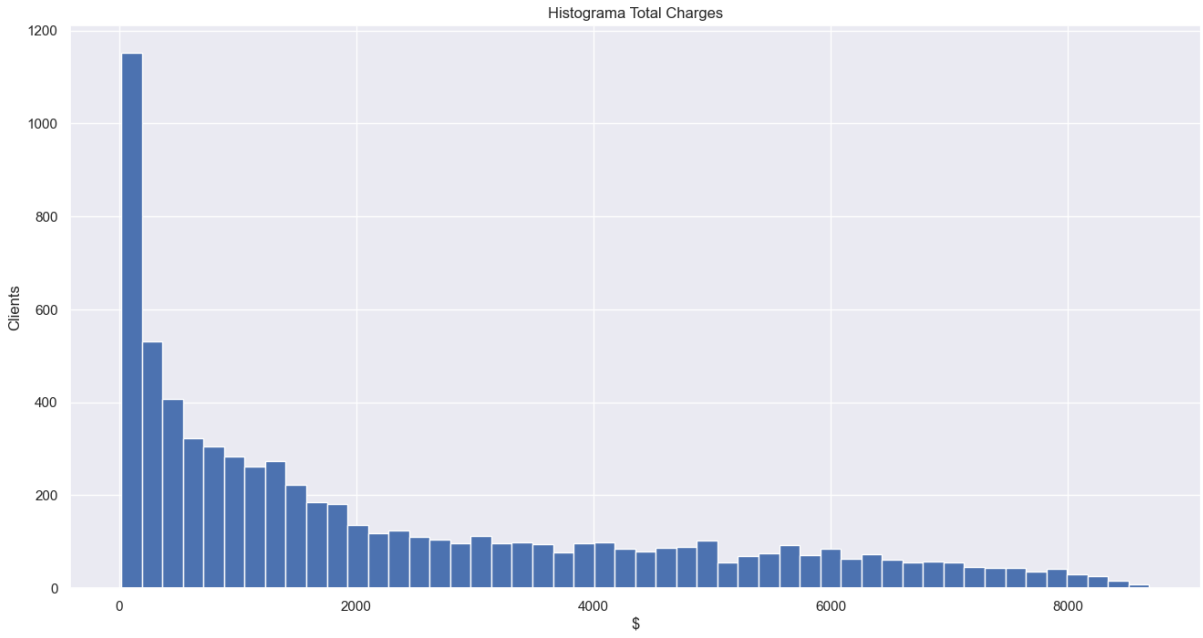

Figura 33: Histograma Total Charge

Sembla que segueix una distribució exponencial.

Total Refunds

	Total Refunds
count	7043.000
mean	1.962
std	7.903
min	0.000
25%	0.000
50%	0.000
75%	0.000
max	49.790

Figura 34: Descripció Total Refunds

No hi ha missings. La majoria de clients no han tingut cap refund. El rang va de 0 a 50.

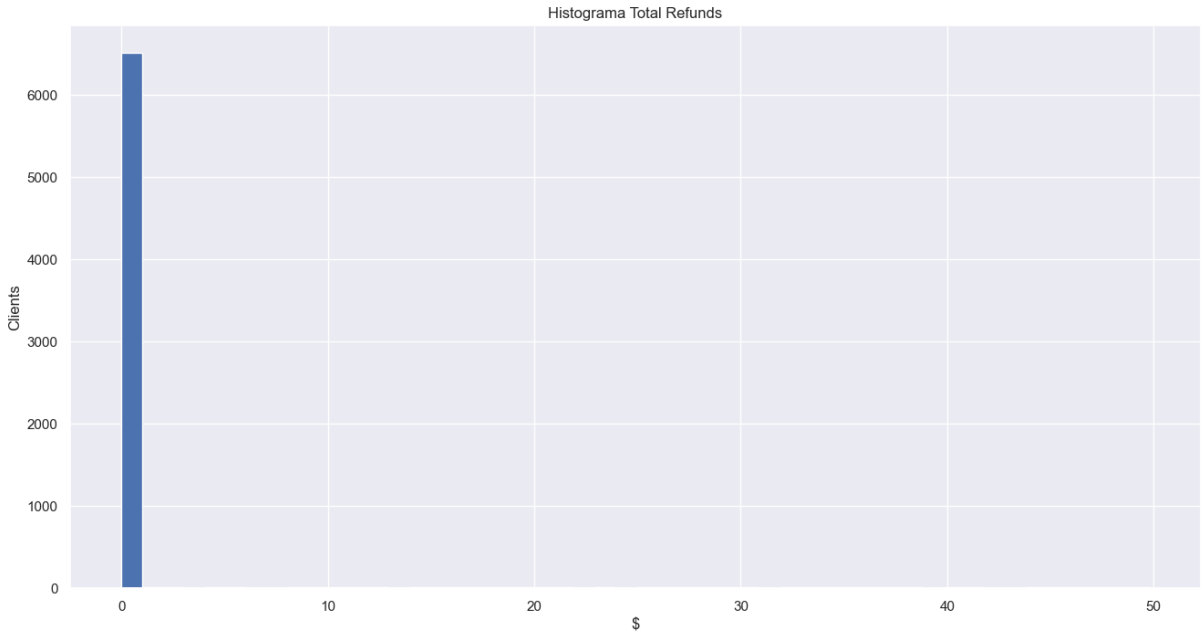

Figura 35: Histograma Total Refunds

És una exponencial tan pronunciada que no es veu res a part dels clients amb 0 de devolució.

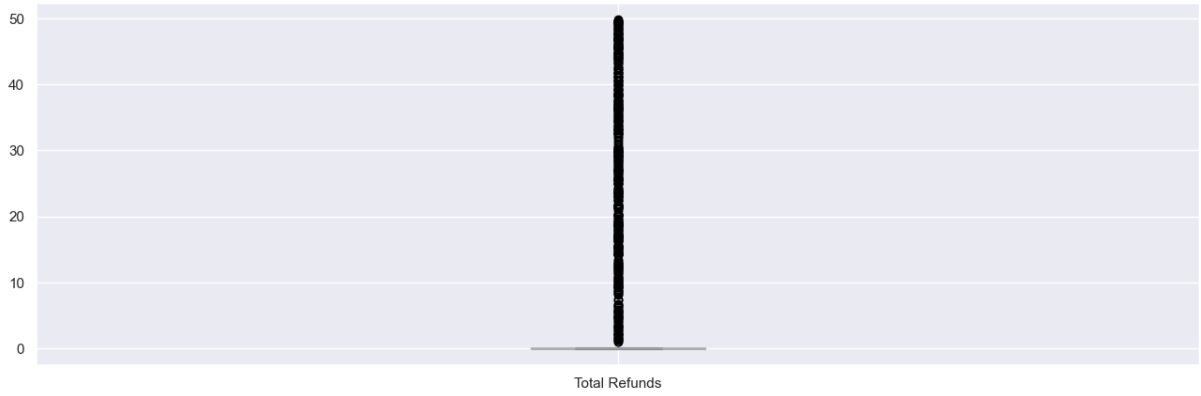

Figura 36: Histograma Total Refunds

Aquí es pot veure millor que si hi ha clients amb devolucions. Decidim modificar la variable ja que al ser un total va fortament correlacionada amb Tenure in Months. No obstant, la variable continua sent molt exponencial on la majoria d'observacions estan al 0. Decidim binaritzar-la. 1 pels que han rebut algun refund, 0 per la resta.

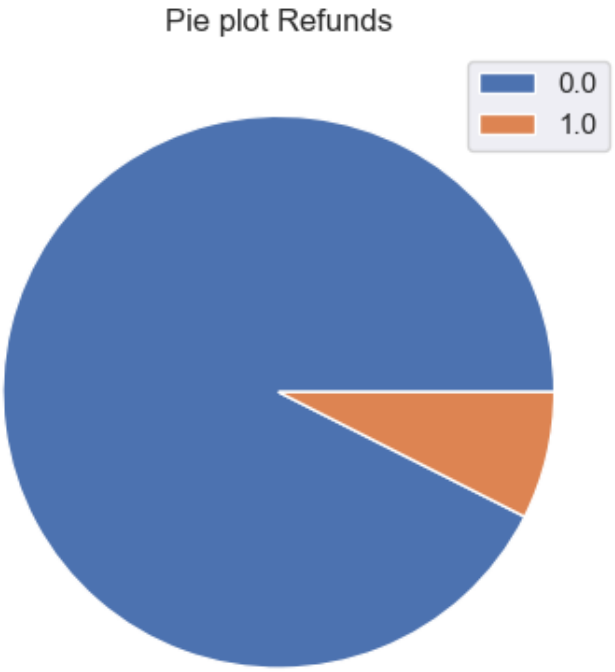

Figura 37: Pie plot Refunds

Total Extra Data Charges

	Total Extra Data Charges
count	7043.000
mean	6.861
std	25.105
min	0.000
25%	0.000
50%	0.000
75%	0.000
max	150.000

Figura 38: Descripció Total Extra Data Charges

La majoria dels clients no tenen càrrega extra per passar-se del límit de dades contractades. No obstant hi ha clients que es passen fins al punt on el màxim és 150.

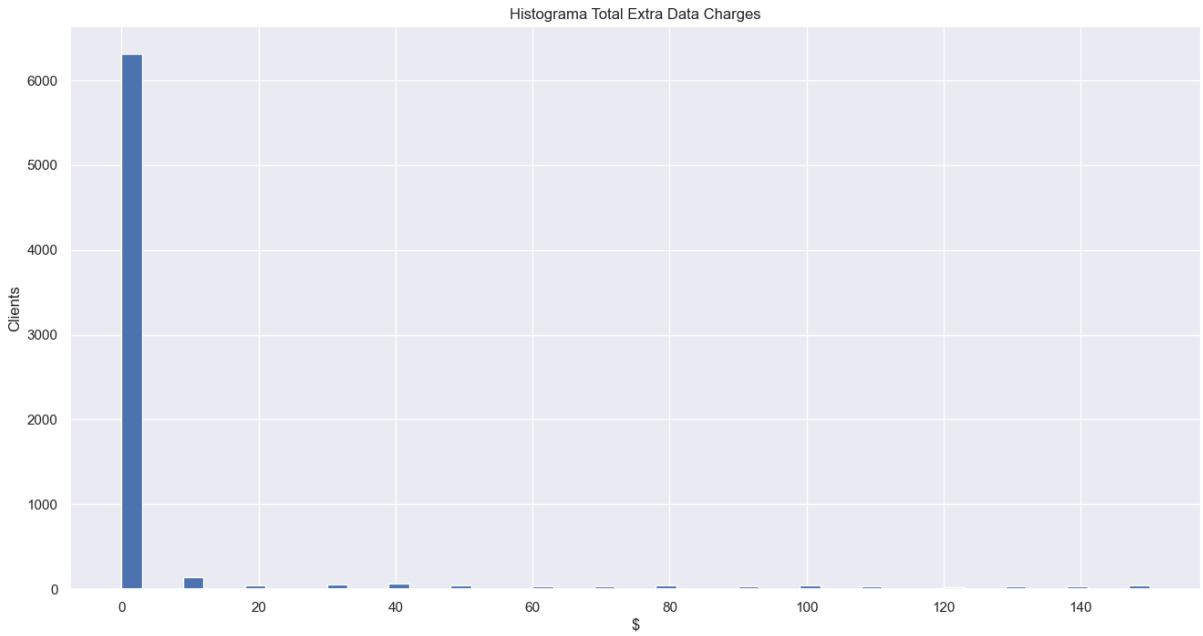

Figura 39: Histograma Total Extra Data Charges

La distribució és exponencial.

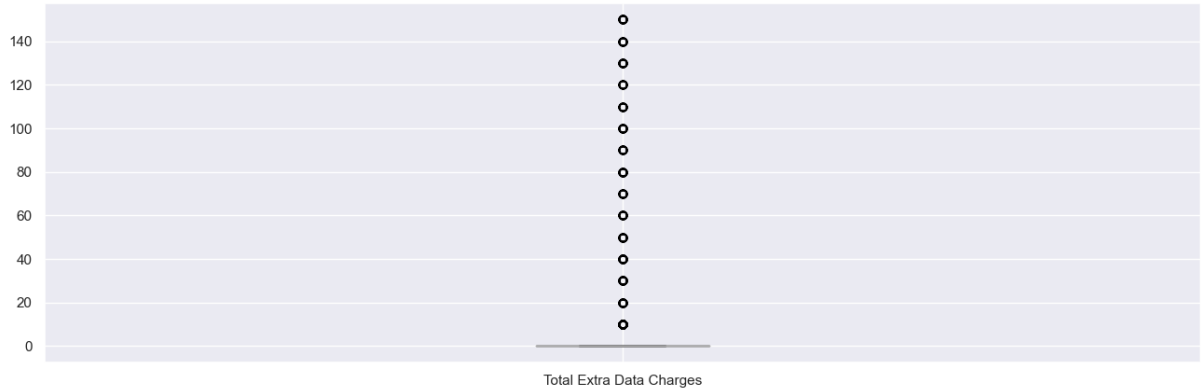

Figura 40: Histograma Total Extra Data Charges

Sembla que quan et passes de les dades contractades, llavors vas a rangs de dades noves disponibles. Com si contractessis més. Per això en l'histograma es veien les barres i en el boxplot els punts estan ben definits i no estan uns mig sobre els altres mig no.

Com és una variable total fem la seva versió average dividint per Tenure in Months.

Total Extra Long Distances Charges

	Total Long Distance Charges
count	7043.000
mean	749.099
std	846.660
min	0.000
25%	70.545
50%	401.440
75%	1191.100
max	3564.720

Figura 41: Descripció Total Extra Long Distances Charges

El rang va de 0 si no tenen càrrecs extra a 3564.72 que un client porta acumulats de càrrecs extra per distància.

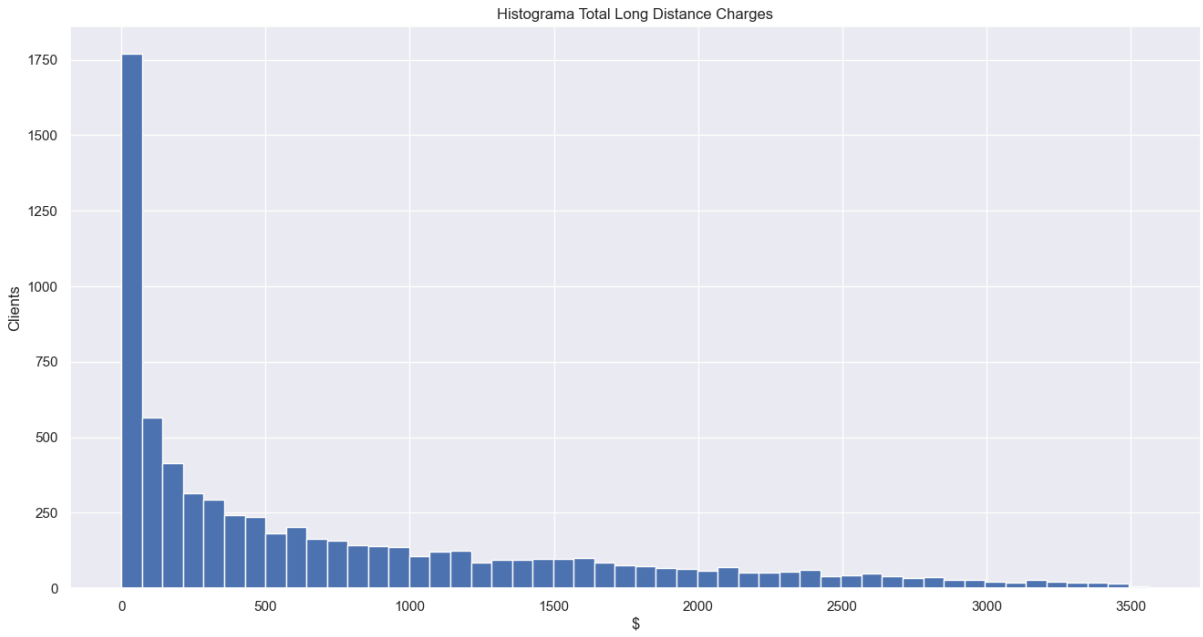

Figura 42: Histograma Total Extra Long Distances Charges

La distribució és exponencial. Com la variable és un total hauríem de fer els seu average. No obstant ja tenim la versió mensual d'aquesta variable així que eliminem Total Extra Long Distances Charges.

Total Revenue

	Total Revenue
count	7043.000
mean	3034.379
std	2865.205
min	21.360
25%	605.610
50%	2108.640
75%	4801.145
max	11979.340

Figura 43: Descripció Total Revenue

No hi ha missings. el mínim és poc més del mínim de Monthly Charge. El màxim quasi arriba als 12k.

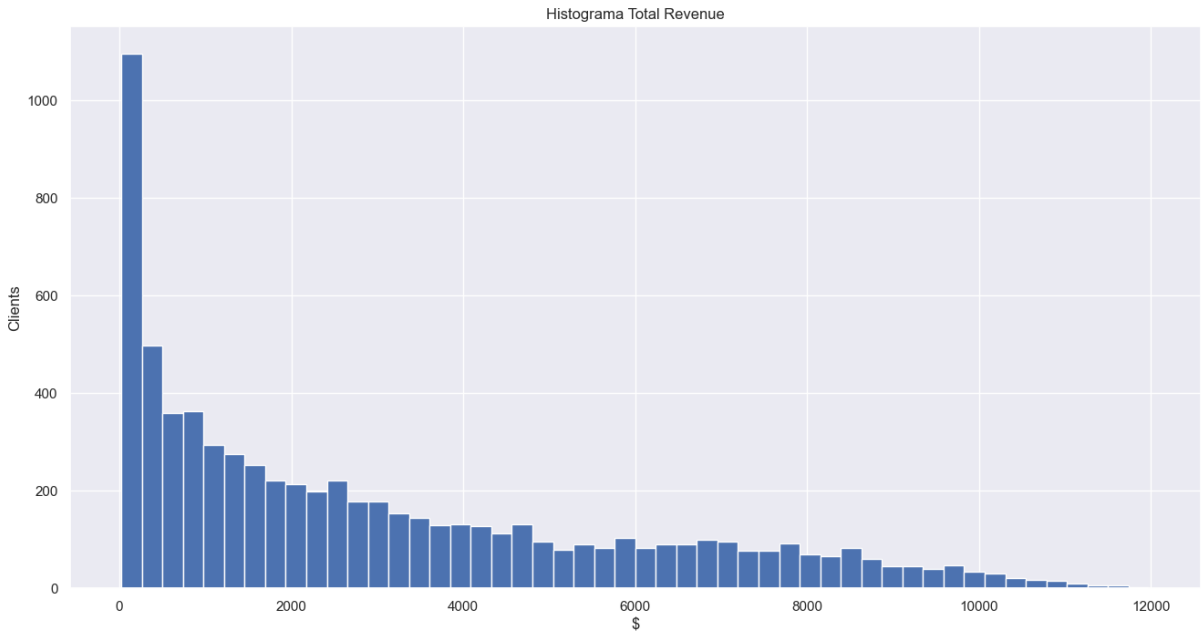

Figura 44: Histograma Total Revenue

La distribució és exponencial com passava amb Total Charges ja que una variable surt de l'altra. Com és una variable total fem la seva versió monthly average.

Online Security

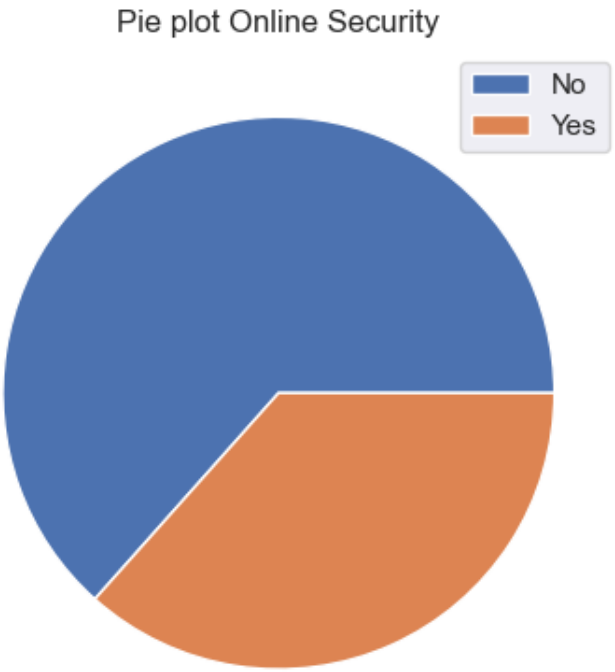

Figura 45: Pie plot Online Security

Hi ha més gent que no té contractat aquest servei que no pas gent que si. Té missings tots els clients que no tenen internet. Els posem com que no tenen contractat aquest seguretat online per imputar.

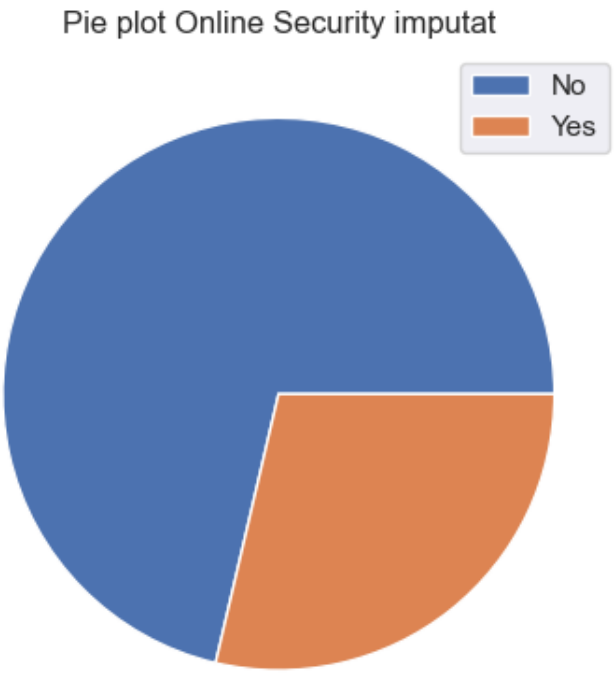

Figura 46: Pie plot Online Security Imputat

Gender

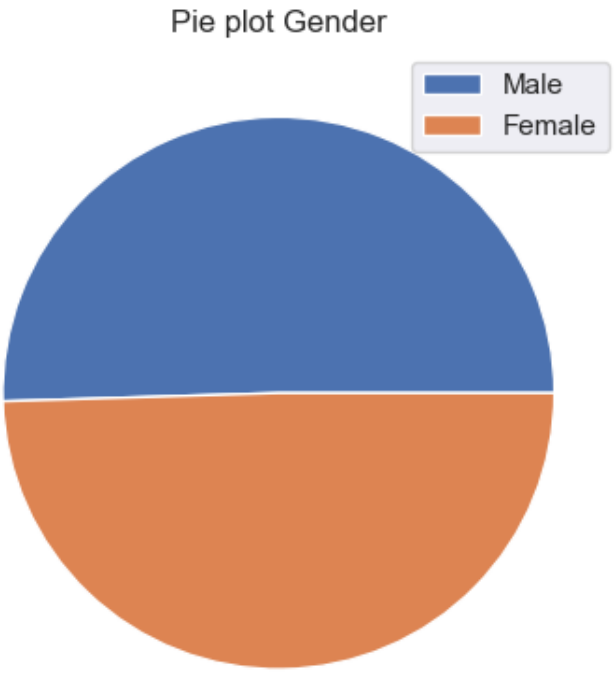

Figura 47: Pie plot Gender

Veiem que el gènere està representat per igual. Al provar de fer un test de chi quadrat hem vist que la variable és independent de la variable resposta i per tant no ens serveix pel nostre model de predicció així que la eliminem. En la següent taula es pot apreciar que el gènere està repartit en diferents categories de Customer Status (només ens interessa Churned i Stayed, però també passa per Joined) equitativament.

Customer Status	Churned	Joined	Stayed
Gender			
Female	939	211	2338
Male	930	243	2382

0.35490117109455394

Figura 48: Taula Gender-Customer Status

Married

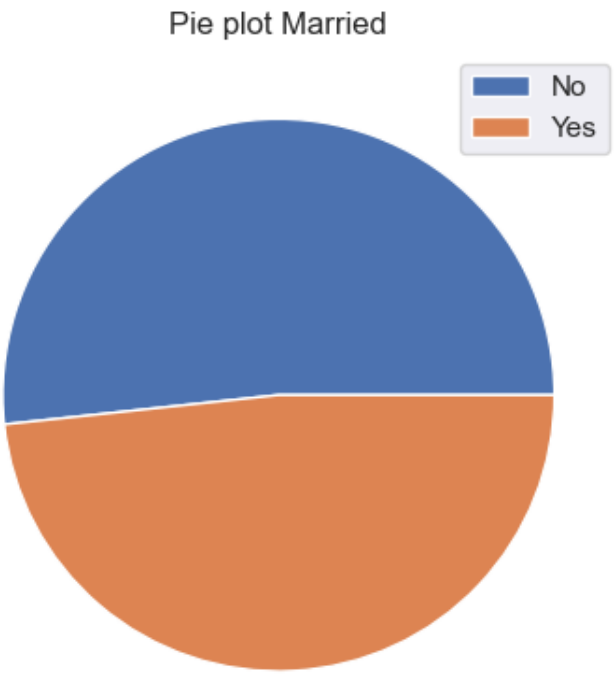

Figura 49: Pie plot Married

Hi ha una lleugera majoria de gent no casada.

Offer

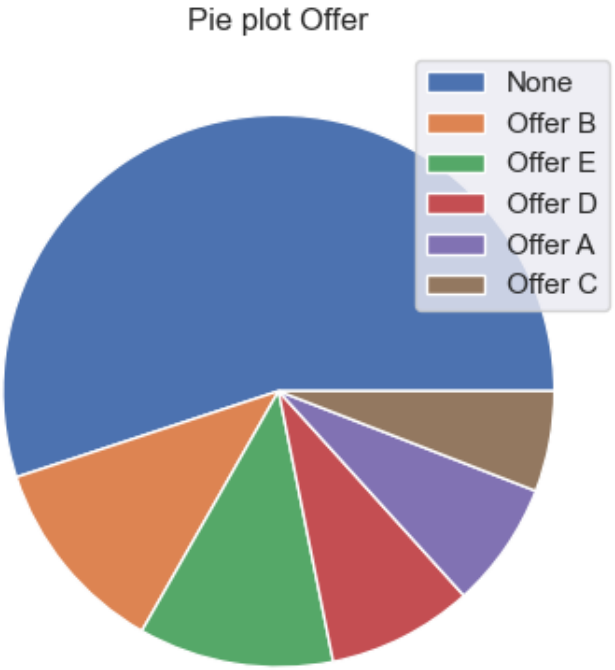

Figura 50: Pie plot Offer

La majoria de clients no està sota cap oferta de màrqueting. La oferta amb més clientela és la B i la que menys és la C.

Phone Service

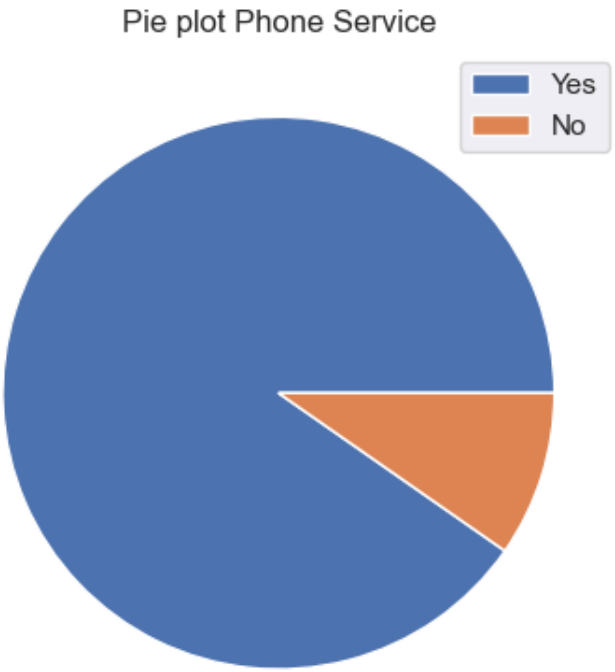

Figura 51: Pie plot Phone Service

Una clara majoria de clients si tenen contractada la línia telefònica de casa.

Multiple Lines

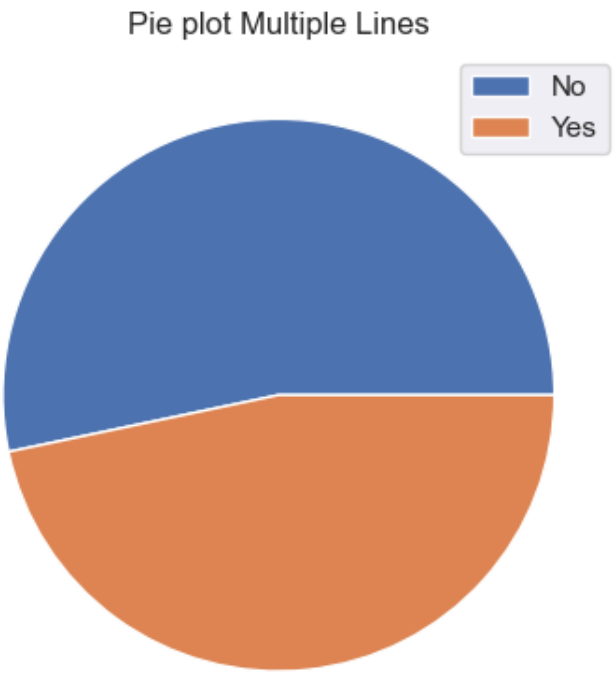

Figura 52: Pie plot Multiple Lines

Sembla que hi ha una lleugera minoria de gent que si té contractades més d'una línia. Però hi ha missing values en aquesta variable. Tots els que no tenen fixe contractat estan com NA i haurien de ser 'No'.

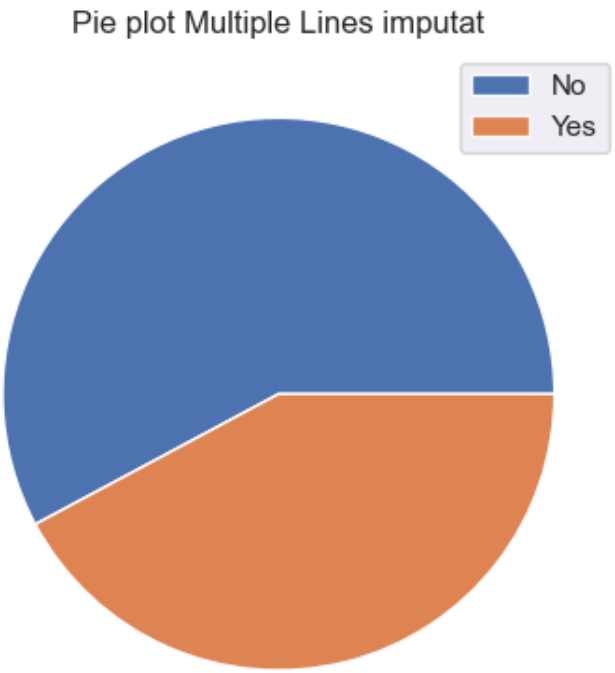

Figura 53: Pie plot Multiple Lines imputat

Ara la majoria del no és una mica més gran.

Internet Service

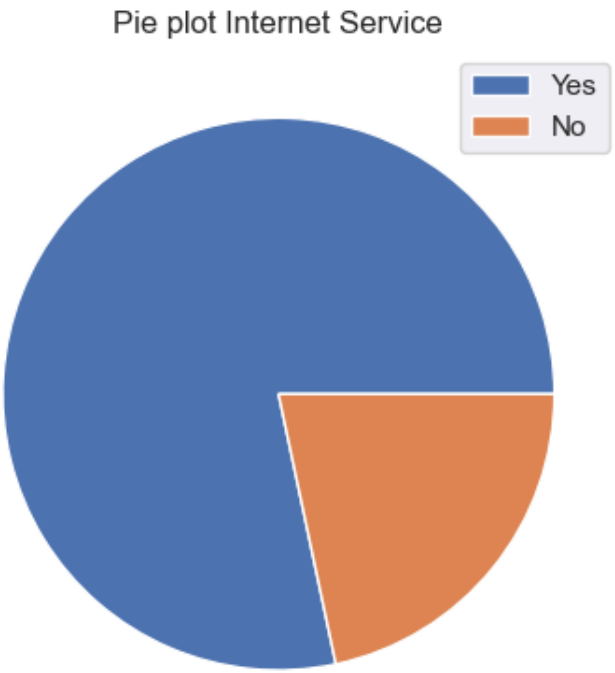

Figura 54: Pie plot Internet Service

Sembla que la major part dels clients si té contractat internet. Com Internet Type ja ens dona la informació de si un client té contractat internet, llavors borrem Internet Service.

Internet Type

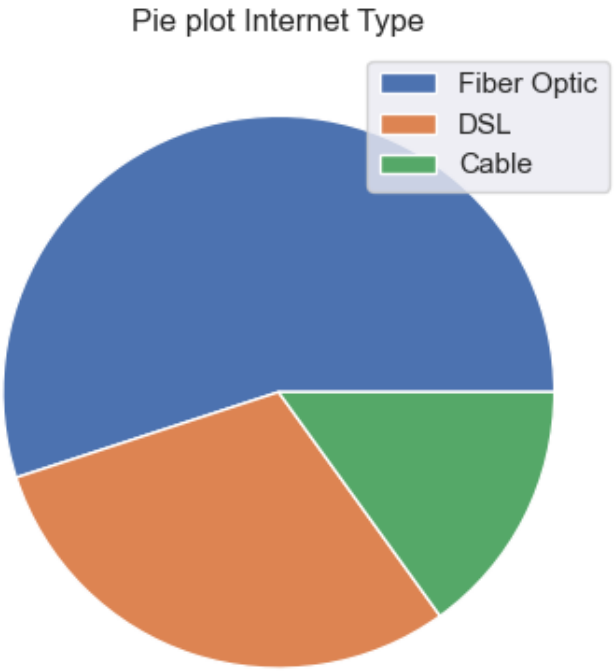

Figura 55: Pie plot Internet Type

El tipus d'internet més contractat és fibra òptica amb més del 50%. Hi ha NAs, tots els clients que no tenen internet contractat no tenen classe assignada. Segons el metadata haurien de ser 'None', així que ho posem.

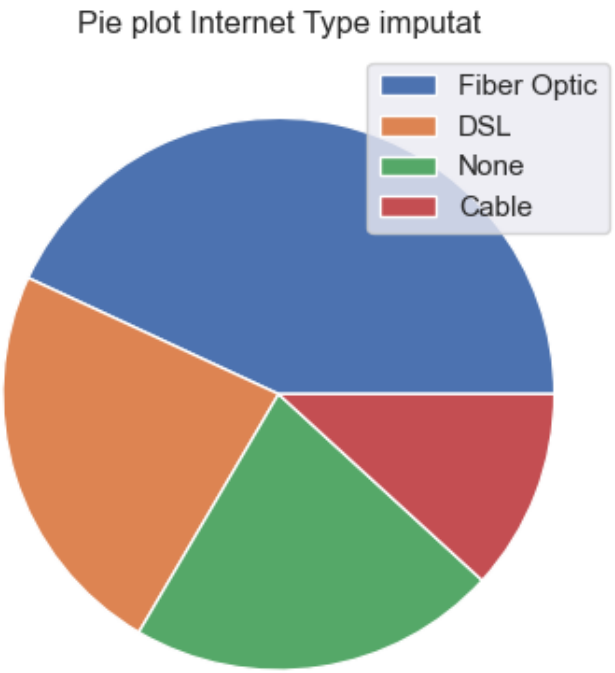

Figura 56: Pie plot Internet Type Imputat

Ara veiem que hi ha 1/4 que no tenen internet. Ja no necessitem Internet Service.

Online Backup

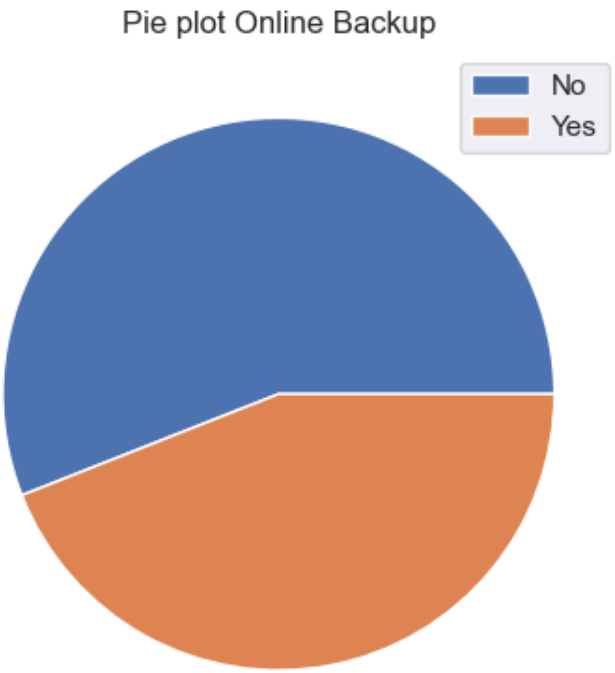

Figura 57: Pie plot Online Backup

Tots els clients sense internet estan com a NA, els posem com a No Online Backup per imputar.

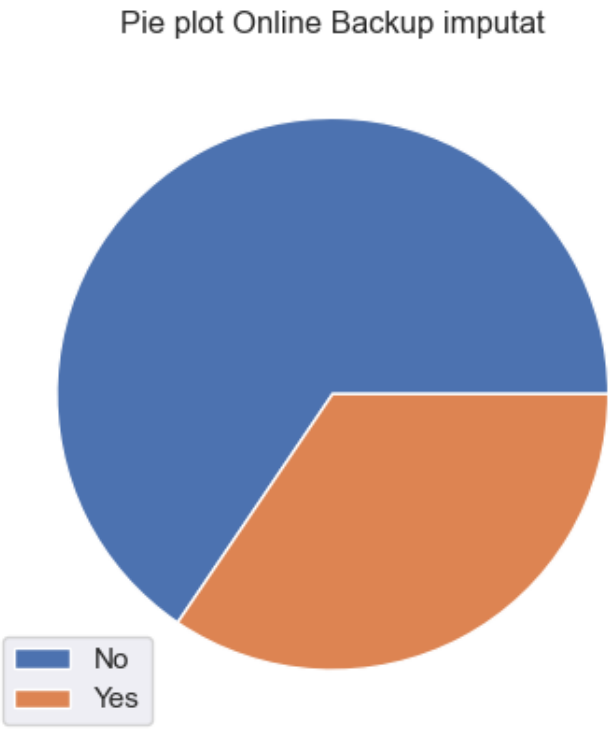

Figura 58: Pie plot Online Backup imputat

Ara es nota més que la majoria de clients no tenen contractat aquest servei.

Device Protection Plan

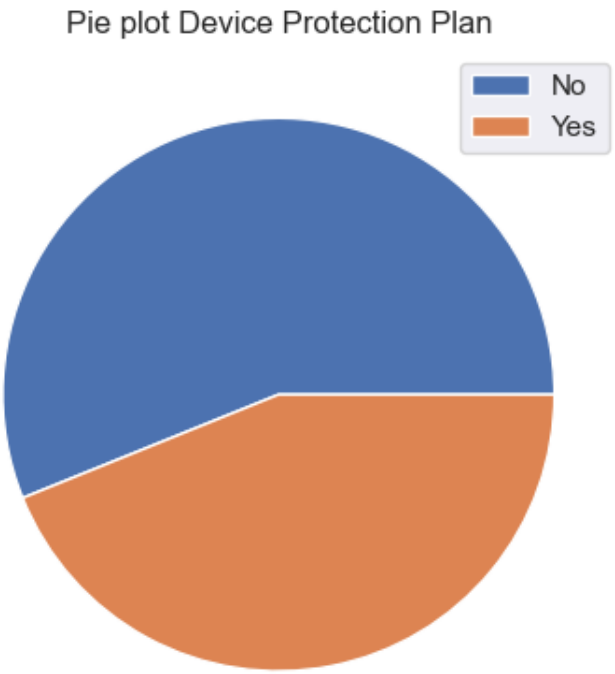

Figura 59: Pie plot Device Protection Plan

Els clients sense internet haurien de ser 'No' i per tant els marquem com a tal, ara eren NA.

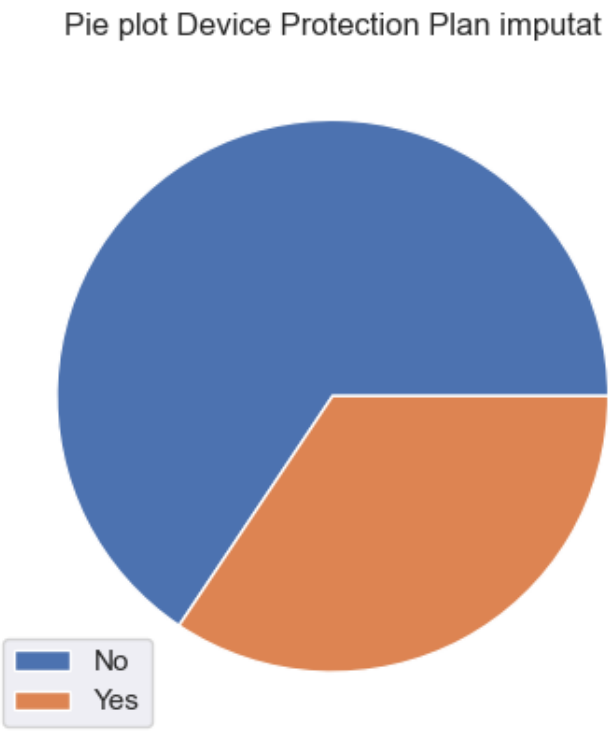

Figura 60: Pie plot Device Protection Plan imputat

Premium Tech Support

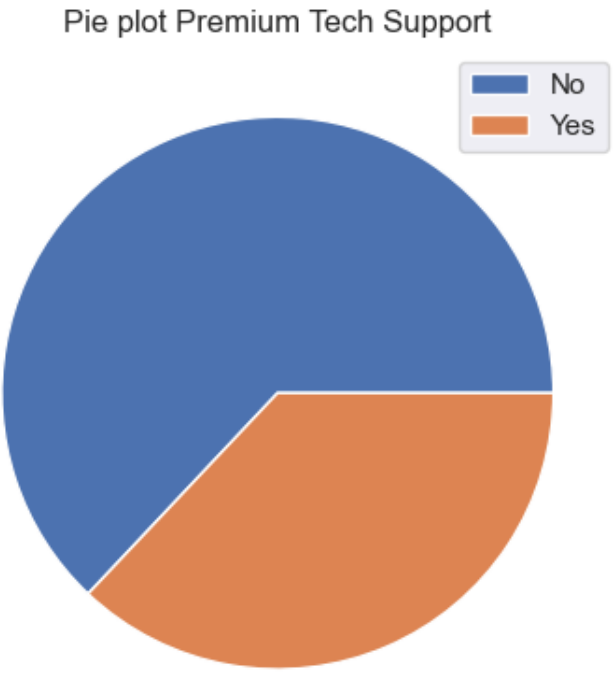

Figura 61: Pie plot Premium Tech Support

Els clients sense internet no tenen valor assignat de la variable. Els posem com que no tenen el servei contractat.

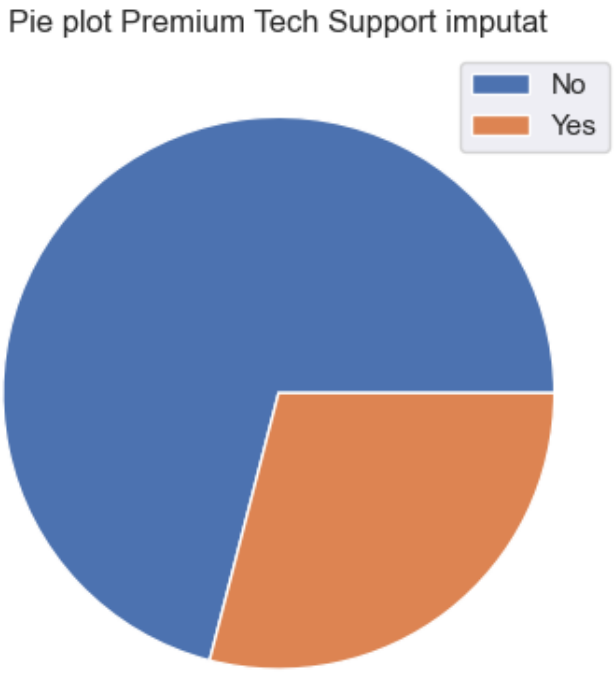

Figura 62: Pie plot Premium Tech Support Imputat

Unlimited Data

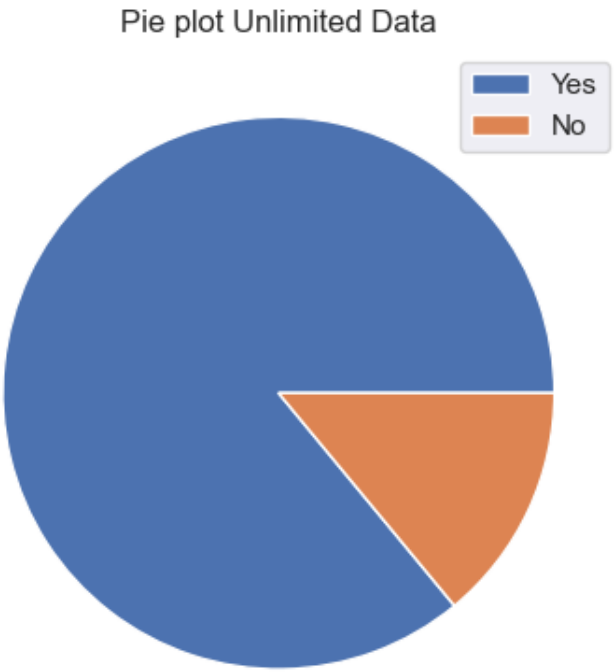

Figura 63: Pie plot Unlimited Data

Aquesta variable té missing values. Tots els clients que no tenen contractat internet, surten com a NA. Els assignem 'No' per imputar.

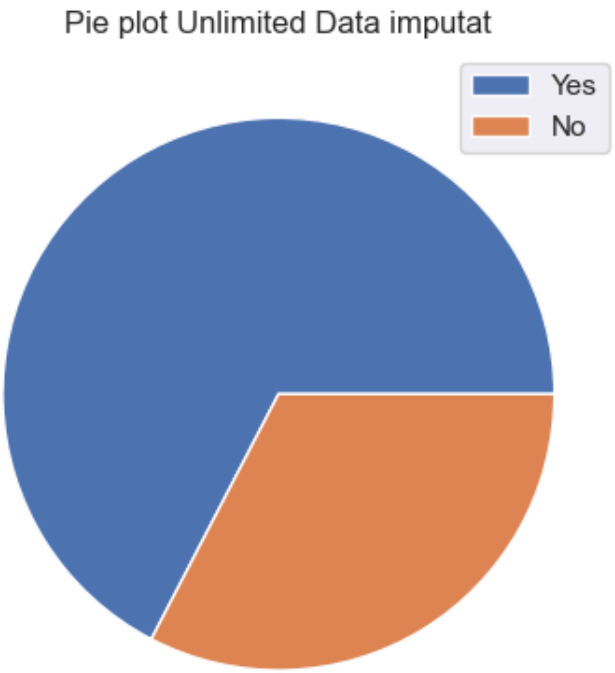

Figura 64: Pie plot Unlimited Data imputat

Contract

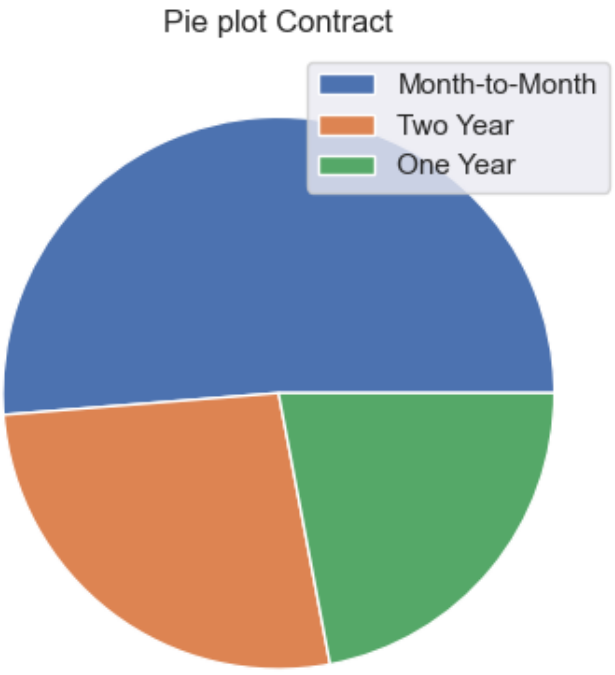

Figura 65: Pie plot Contract

Paperless Billing

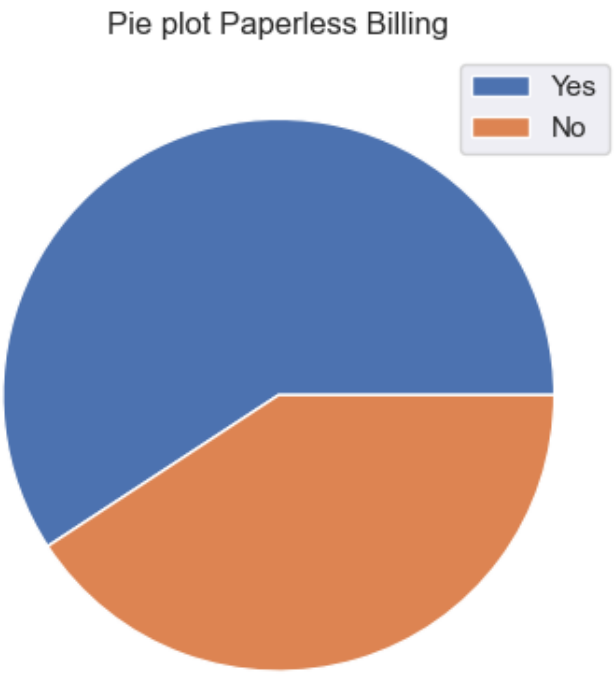

Figura 66: Pie plot Paperless Billing

Payment Method

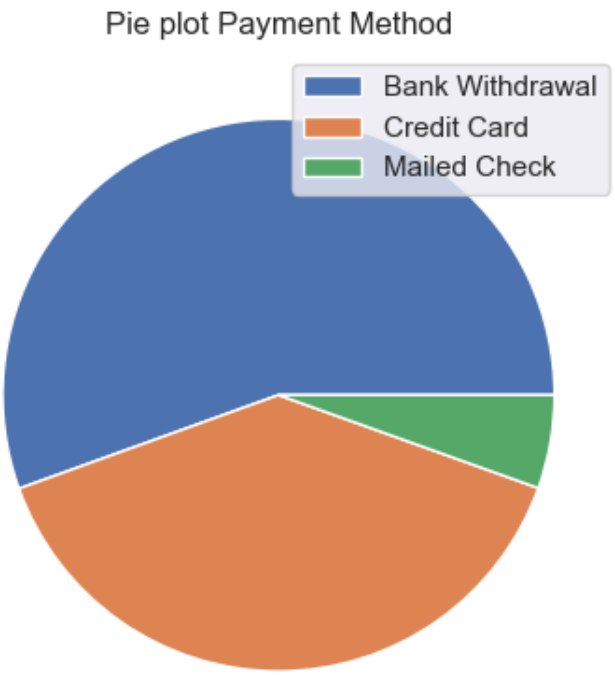

Figura 67: Pie plot Payment Method

Churn Category

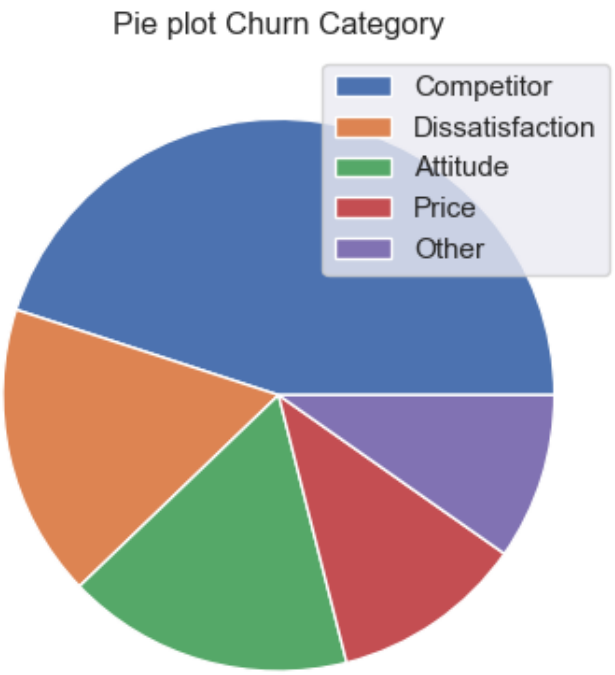

Figura 68: Pie plot Churn Category

La categoria més seleccionada és competitor. Això indica que les altres companyies tenten als clients amb ofertes. Price no és gaire seleccionat però ve a ser el mateix, ja que o estàs en alguna companyia o no tens connexió. Attitude és molt gran i per tant potser s'hauria de revisar els treballadors de cara al públic ja que potser no estan oferint el millor servei possible. Dissatisfaction també és molt seleccionat, així que potser s'ha de revisar els serveis que dona la companyia, per exemple la velocitat de xarxa o si la xarxa té caigudes. Churn Reason ens dona la mateixa informació una mica més detallada, però no ens serveix per treure cap conclusió més.
Per el model de predicció no utilitzem ni Churn Reason ni Churn Category.

Customer Status

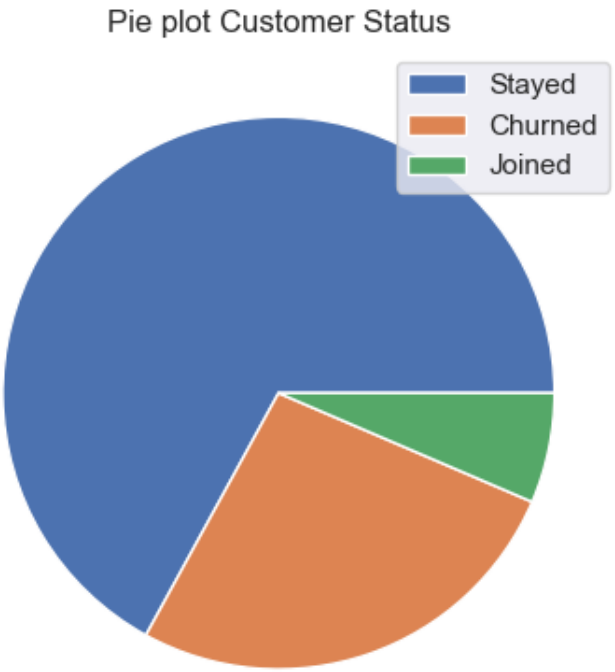

Figura 69: Pie plot Customer Status

Aquesta variable és d'on trèiem si el client marxa o es queda. Com pot ser que els clients nous només durin 1 mes i si són nous és difícil que contractin tots els serveis addicionals, per evitar ajuntar-los amb els clients antics, només utilitzarem Stayed i Churned per fer la predicció. No obstant, per clustering utilitzarem totes les dades.

Streaming TV

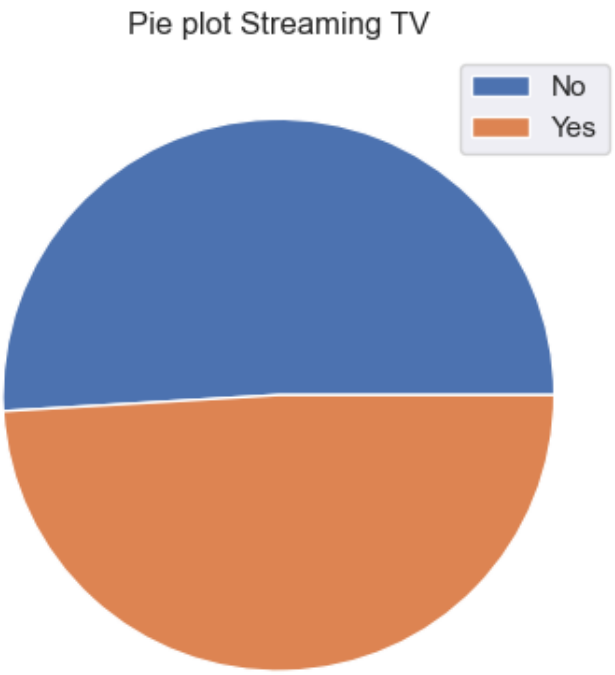

Figura 70: Pie plot Streaming TV

Tant aquesta variable com les altres dues de streaming passa que si el client no té internet contractat, les variables de streaming surten com a missing value. No obstant, segons la metadata haurien de ser 'No'. Així que li assignem el valor que diu la metadata.

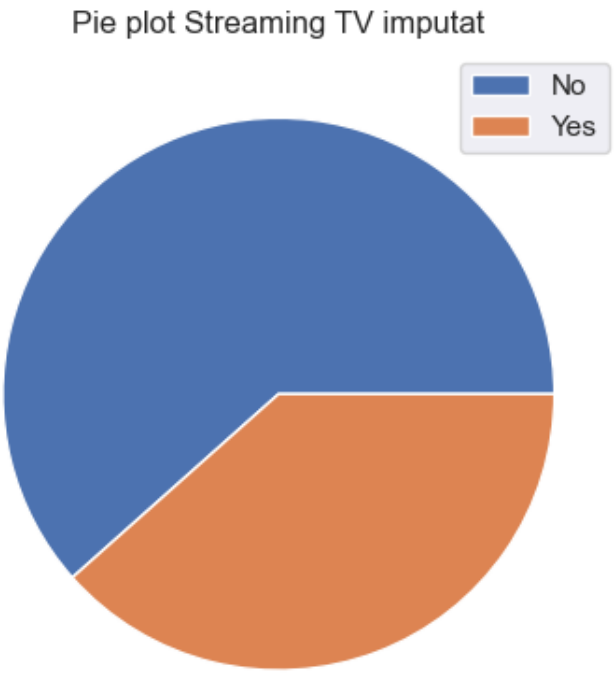

Figura 71: Pie plot Streaming TV imputat

Streaming Music

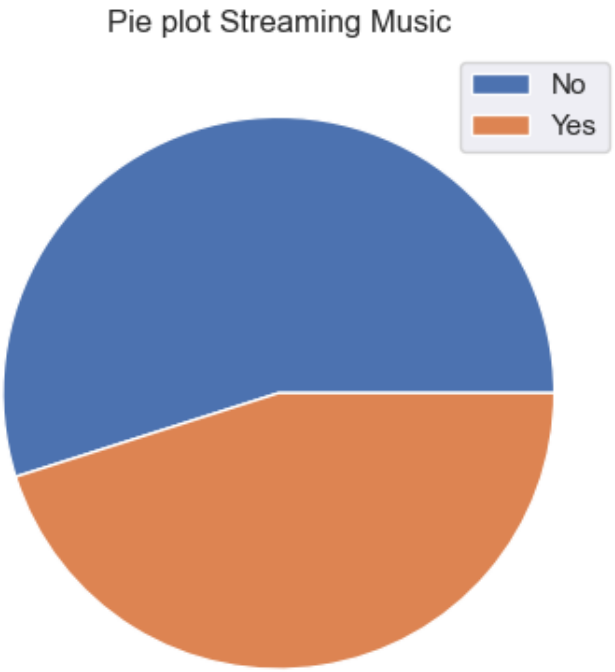

Figura 72: Pie plot Streaming Music

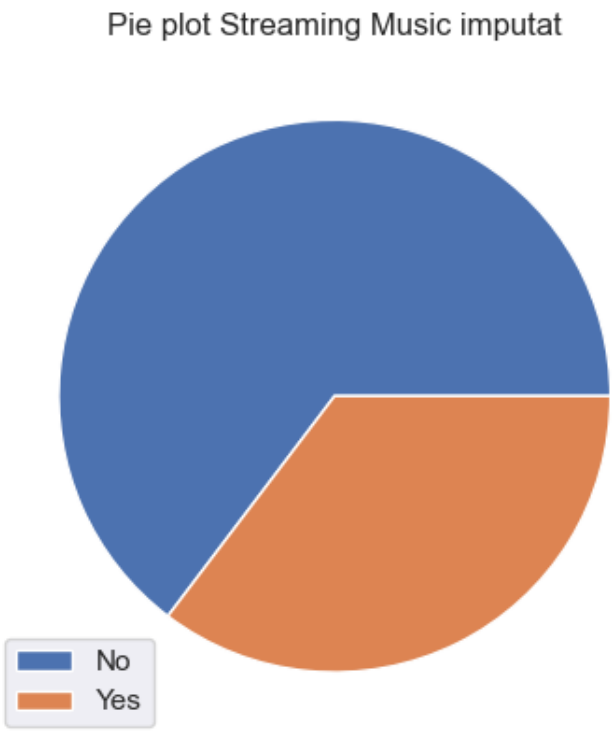

Figura 73: Pie plot Streaming Music imputat

Streaming Movies

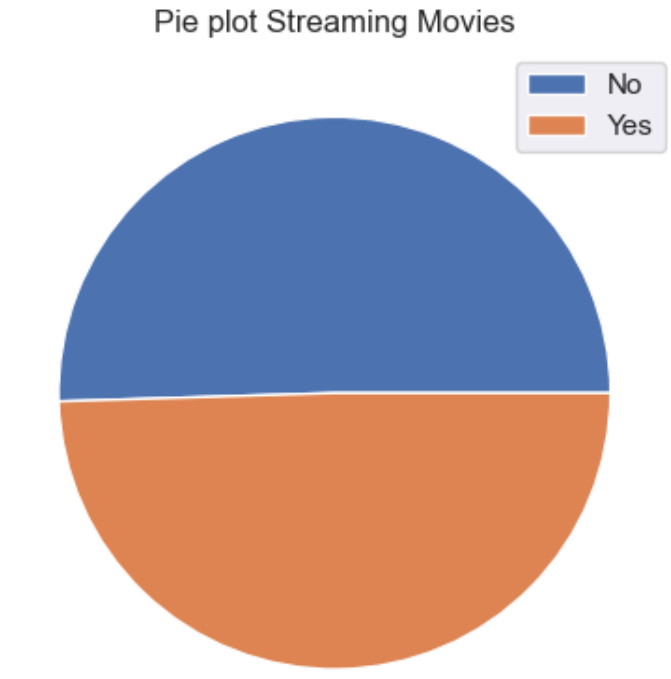

Figura 74: Pie plot Streaming Movies

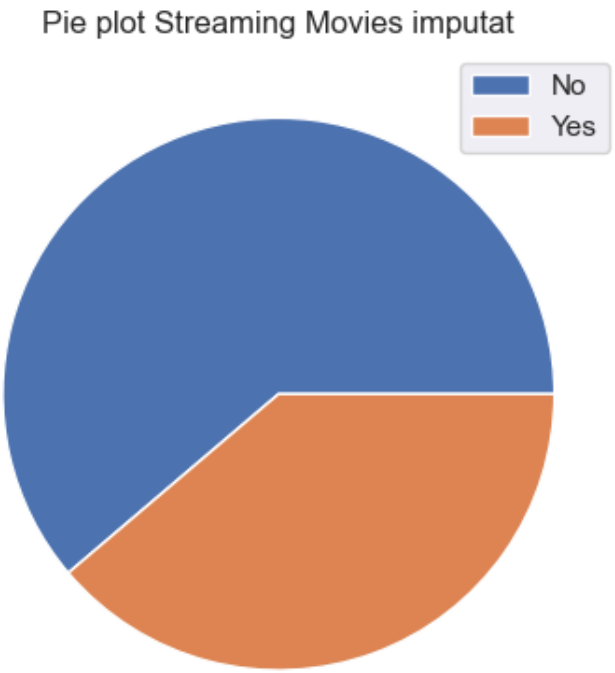

Figura 75: Pie plot Streaming Movies imputat

3.3 Recodificació

Hem vist que hi ha moltes variables categòriques. La majoria binàries. Hi ha variables que van relacionades, com podria ser els serveis de streaming o serveis premium que pots contractar si tens internet contractat. El que podem fer és ajuntar aquestes variables en una sola que sigui un count de quants serveis té aquell client.

3.3.1 Streaming

Streaming Music	No	Yes
Streaming Movies		
No	4180	131
Yes	375	2357

```
0.0
```

Figura 76: Exemple Taula contingència Streaming

Entre les variables de streaming si fem tests de chi quadrat ens dona un p-valor de 0. Si la nostra alfa (llindar de decisió per la confiança) és del 0.05. $0<0.05$ i per tant rebutgem la hipòtesi nul·la del test, que és que les variables són independents. Per tant sabem que estan relacionades. En la taula de contingència d'exemple també es veu clar que la categoria de una influeix en l'altra, en les altres combinacions també passa. Així que ajuntem les tres variables un una variable 'Total Streaming' que és un count de serveis de streaming.

3.3.2 Premium

En aquest cas farem una cosa semblant. Fem el mateix procediment. Provem les variables que sembli que pugui estar relacionades i fem test chi quadrat i taules de contingència. Al final hem decidit ajuntar: 'Premium Tech Support', 'Multiple Lines', 'Online Backup', 'Device Protection Plan', 'Online Security' i 'Unlimited Data'. Totes relacionades amb tenir internet.

3.4 Anàlisi Correlacions 2

Ara tenim noves variables així que hem de tornar a fer un anàlisi de correlacions.

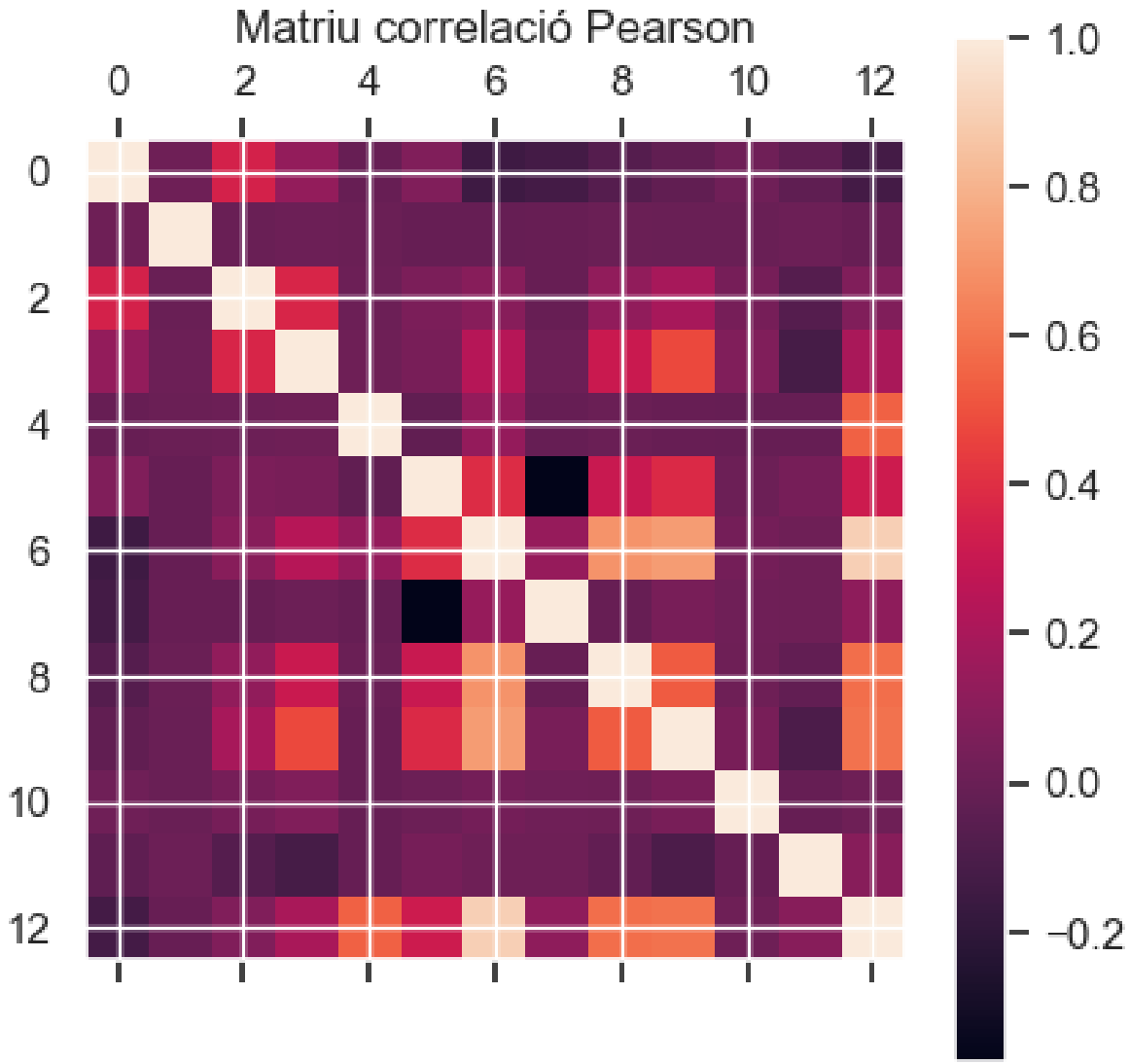

Figura 77: Matriu de Correlacions amb les noves variables

Els índexs per les variables són:
0: Number of Dependents, 1: Zip Code, 2: Number of Referrals, 3: Tenure in Months, 4: Avg Monthly Long Distance Charges, 5: Avg Monthly GB Download, 6: Monthly Charge, 7: Age_disc, 8: Total Streaming, 9: Premium Services, 10: Refunds, 11: Average Monthly Extra Data Charges, 12: Average Monthly Revenue.

Podem veure que les correlacions altes són les següents:
Average Monthly Revenue - Monthly Charge: Average Monthly Revenue al ser una combinació d'altres

variables està correlacionada amb més d'una. La podem eliminar ja que supera el nostre límit de 0.8 amb Monthly Charge.

Premium Services & Total Streaming: Aquestes variables estan correlacionades fortament amb Monthly Charge, però no superen el nostre llindar, així que les mantenim.

3.5 Tests Independència

Numèriques

Fem tests de Kruskal-Wallis amb totes les variables contra la variable resposta. Les dues variables que tenen un p-valor per sobre de 0.05 i que per tant considerem independents de la variable resposta són Zip Code i Avg Monthly Long Distance Charges. Com Zip Code és bastant especial la deixem.

Pel que fa a Avg Monthly Long Distance Charges fem un boxplot múltiple per veure si es solapen.

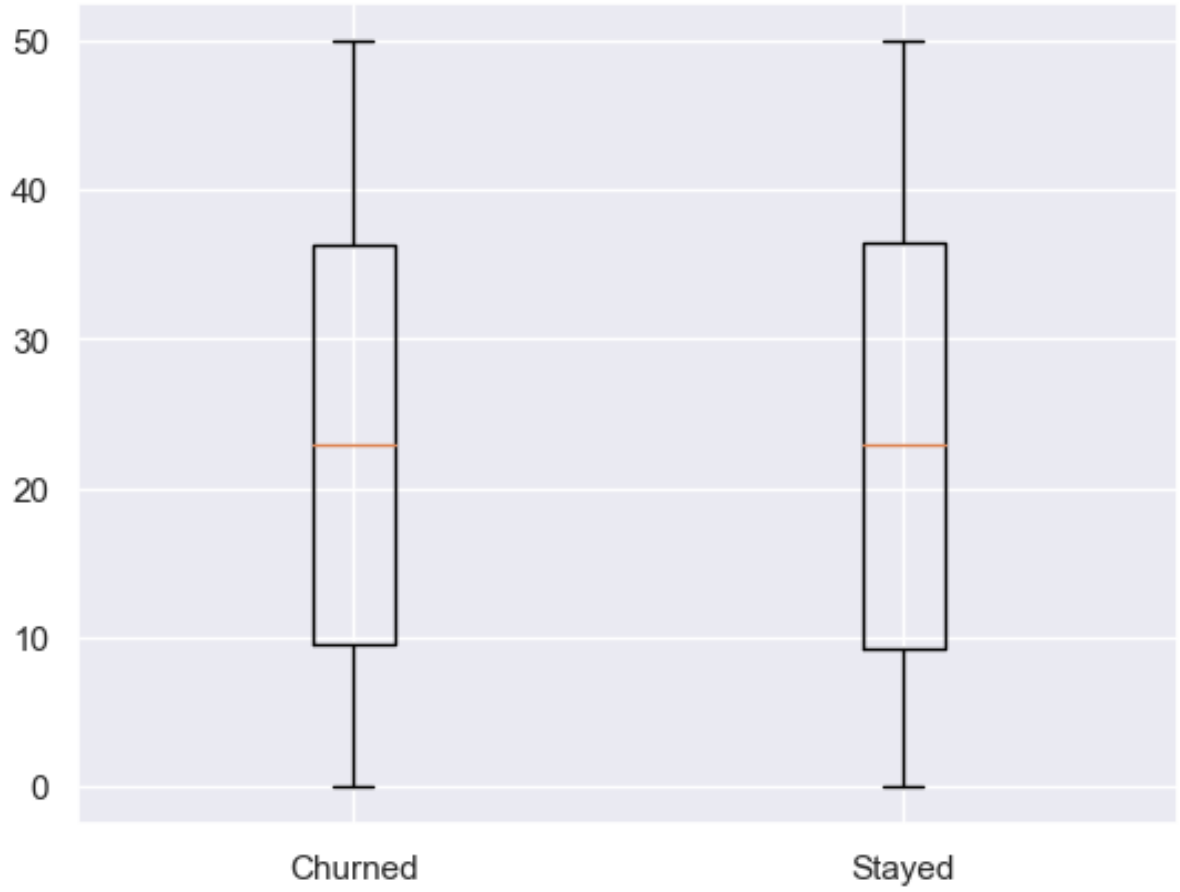

Figura 78: Boxplot múltiple de Avg Monthly Long Distance Charges segons la modalitat de Churn Status

Efectivament els boxplots de les dues modalitats de la variable resposta són iguals, així que eliminem Avg Monthly Long Distance Charges.

Categòriques

La variable que no trenca la hipòtesi nul·la de independència és Phone Service. Si visualitzem la taula de contingència veiem que Phone Service és totalment independent de Customer Status. Eliminem Phone Service.

Customer Status	Churned	Stayed
Phone Service		
No	170	474
Yes	1699	4246

0.26256666476273127

Figura 79: Taula de Contingència Phone Service - Customer Status

3.6 Anàlisi de riscos i biaixos de les variables

Les variables restants no sembla que tinguin cap risc ni biaix potencial així que seguim preparant les dades.

4 User Profiling

Volem saber característiques que diferencien als clients que marxen i als clients que es queden, per això utilitzarem la variable Customer Status; els clients que marxen són els churned i els que es queden són els stayed. No considerem els joined ja que ens posen soroll a les dades dels clients, ja que no saps si marxaran o no i com acaben d'unir-se no tenen gaires serveis contractats.
La estratègia és mirar mitjanes i histogrames per les variables numèriques i mirar proporcions per les variables categòriques.

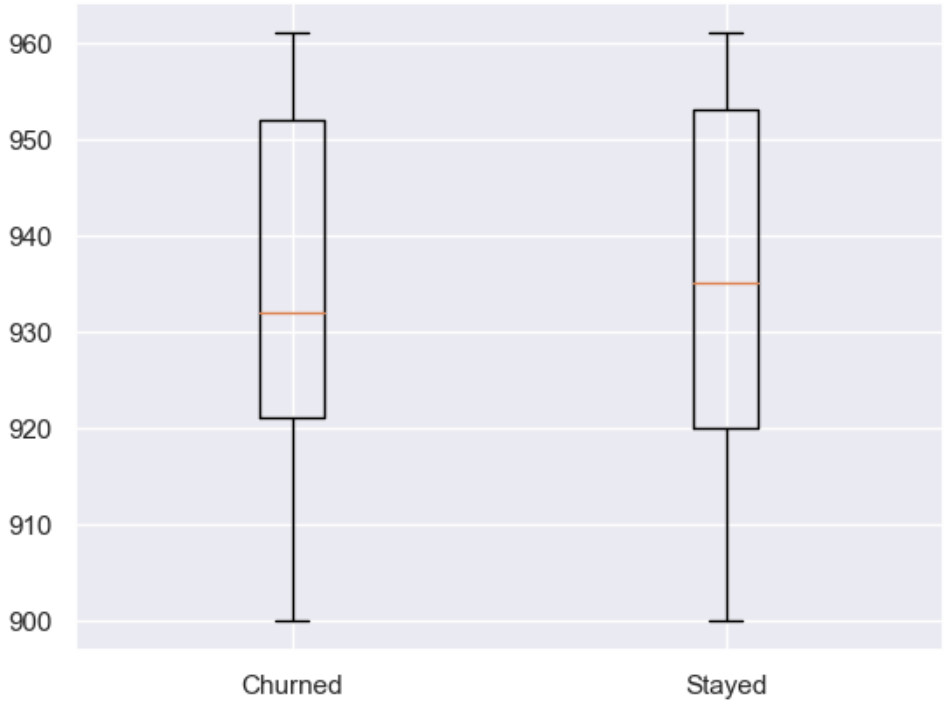

Figura 80: Zip Code Churned vs Zip Code Stayed

Anteriorment ja hem dit que el Zip Code es distribuïa independentment de Customer Status. En aquest multiple boxplot es pot apreciar ja que es solapen i la mitjana és la mateixa.

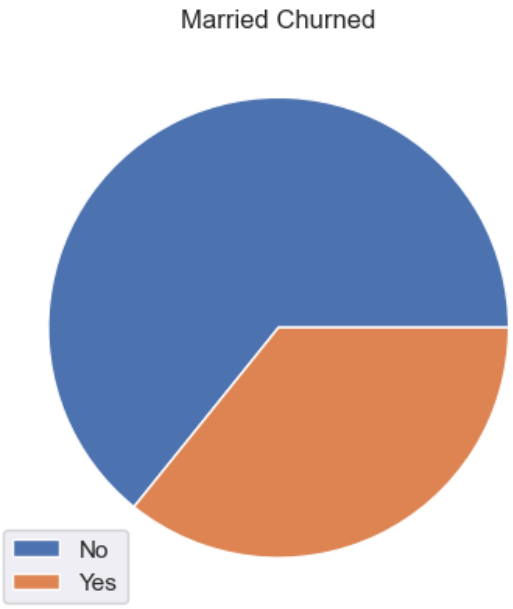

Figura 81: Married Churned

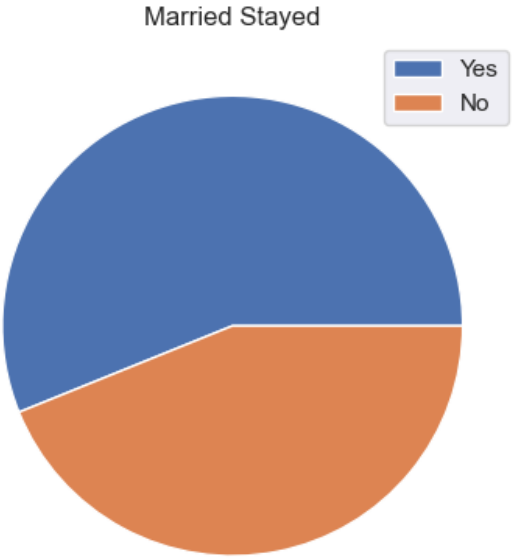

Figura 82: Married Stayed

La proporció de clients casats és més gran entre els clients que es queden.

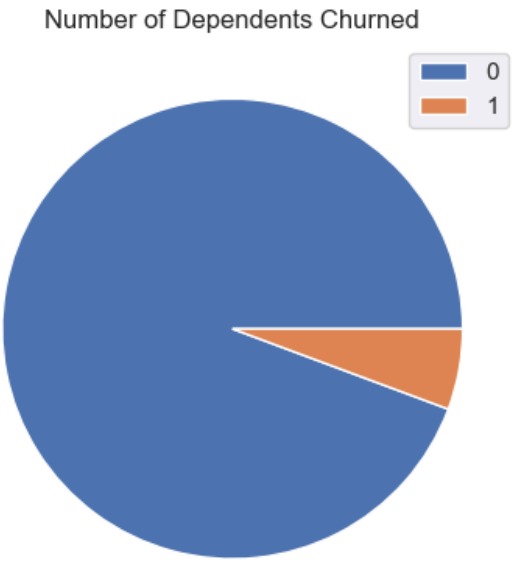

Figura 83: Dependents Churned

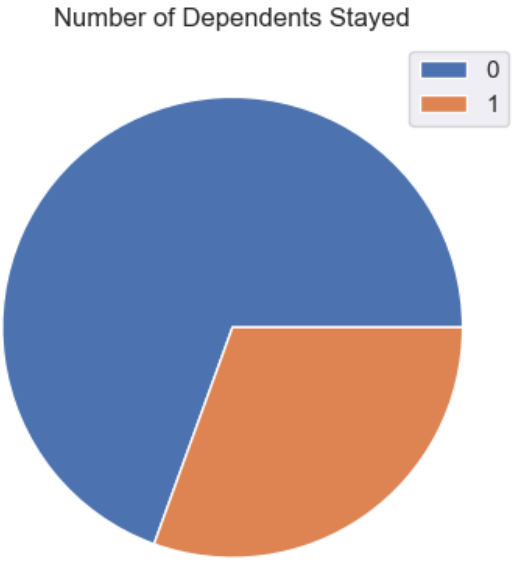

Figura 84: Dependents Stayed

La proporció de clients amb dependents és més alta entre els clients que es queden.

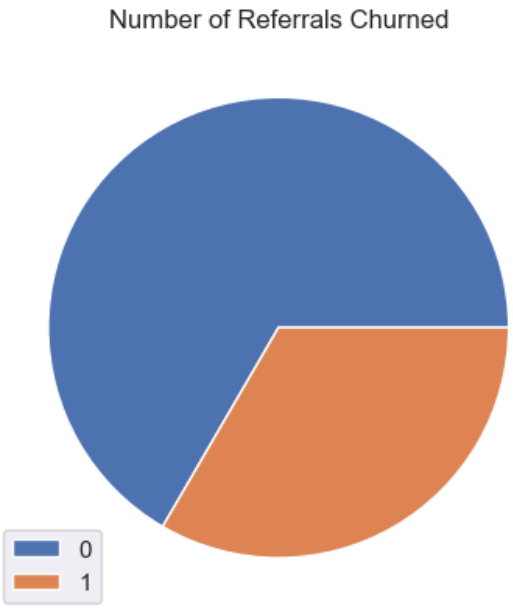

Figura 85: Referrals Churned

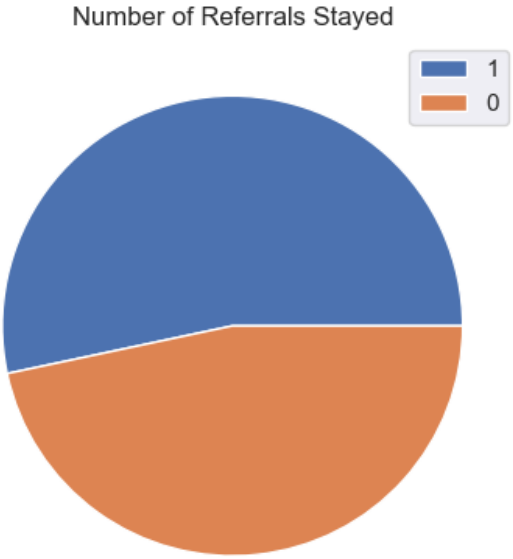

Figura 86: Referrals Stayed

La proporció de clients que ha referit a un conegut o més és més gran entre el clients que s'han quedat en la companyia.

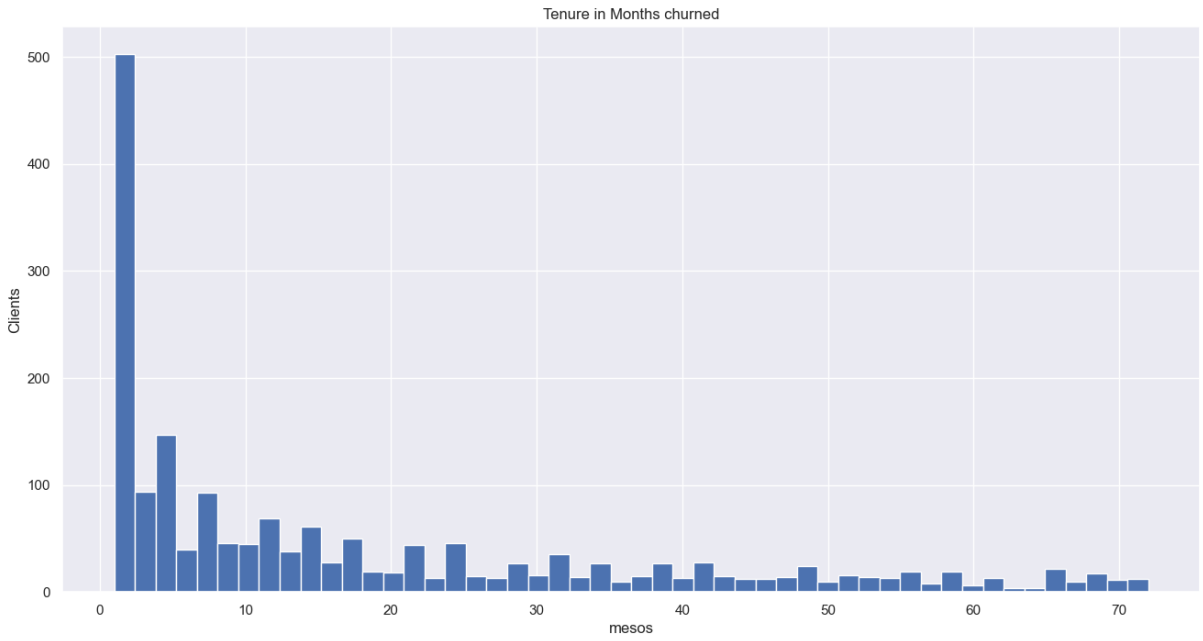

Figura 87: Tenure Churned

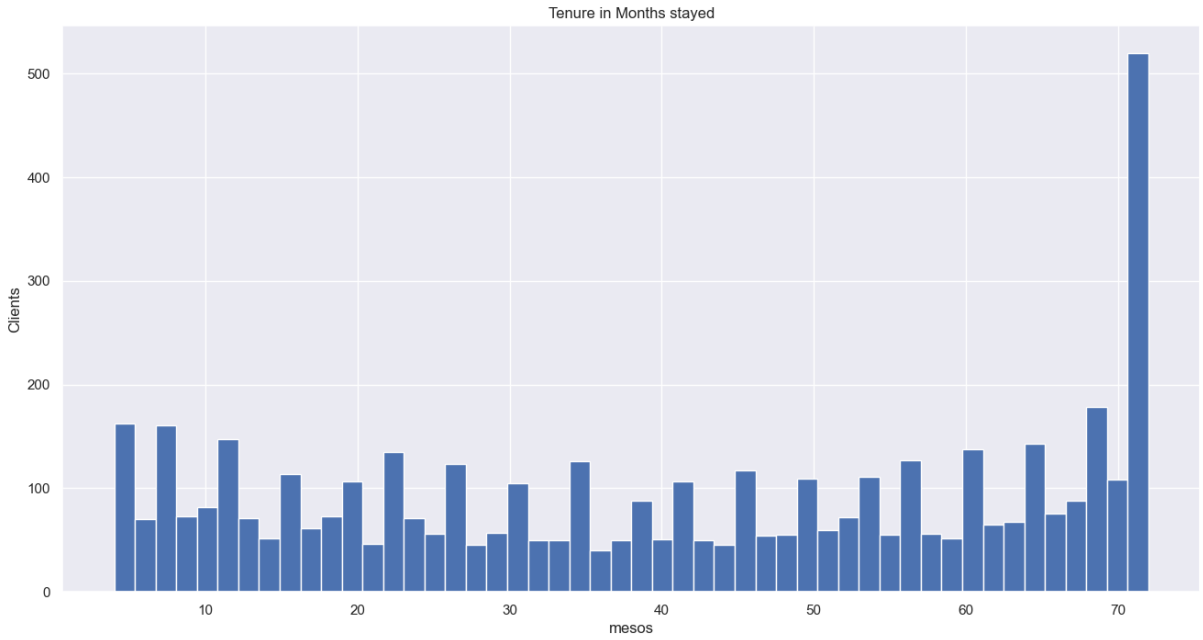

Figura 88: Tenure Stayed

Els clients que porten menys temps es troben en els que marxen mentre que els que porten més temps es queden.

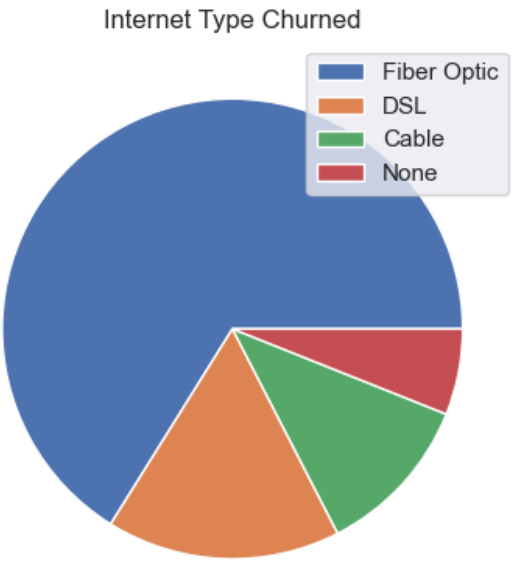

Figura 89: Internet Type Churned

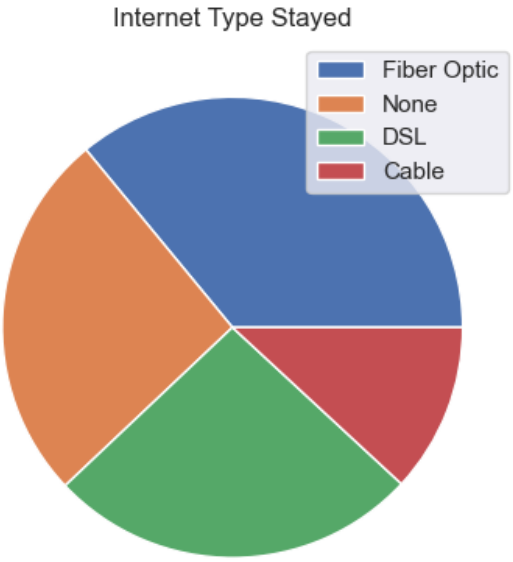

Figura 90: Internet Type Stayed

La majoria de clients que marxen tenen contractada fibra òptica.

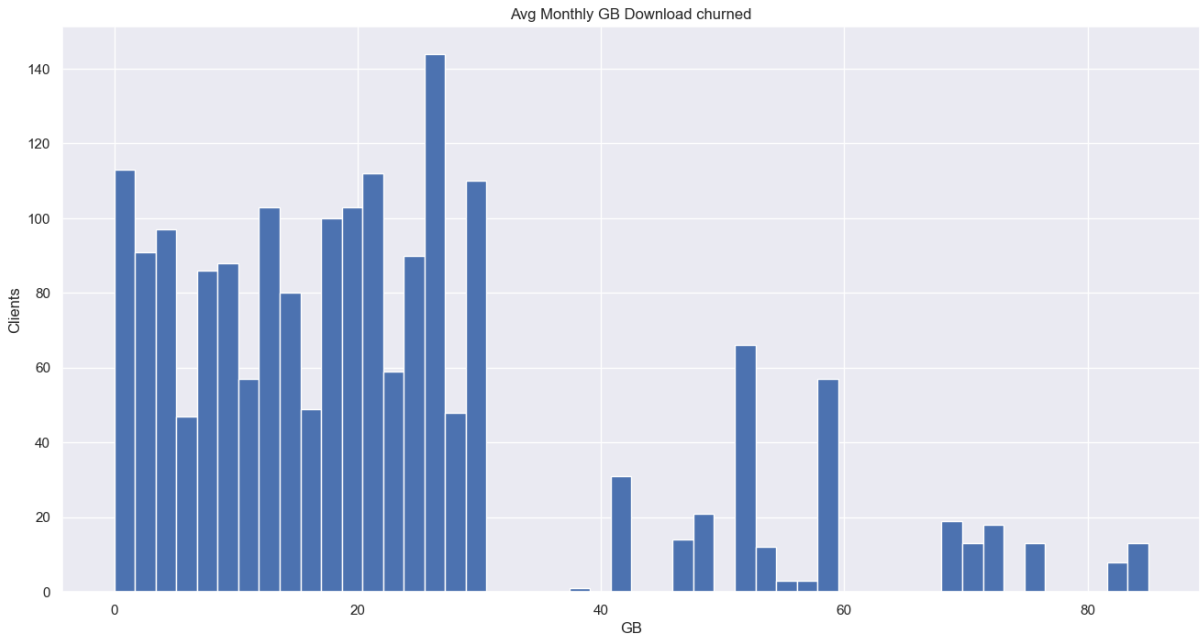

Figura 91: Avg Monthly GB Download Churned

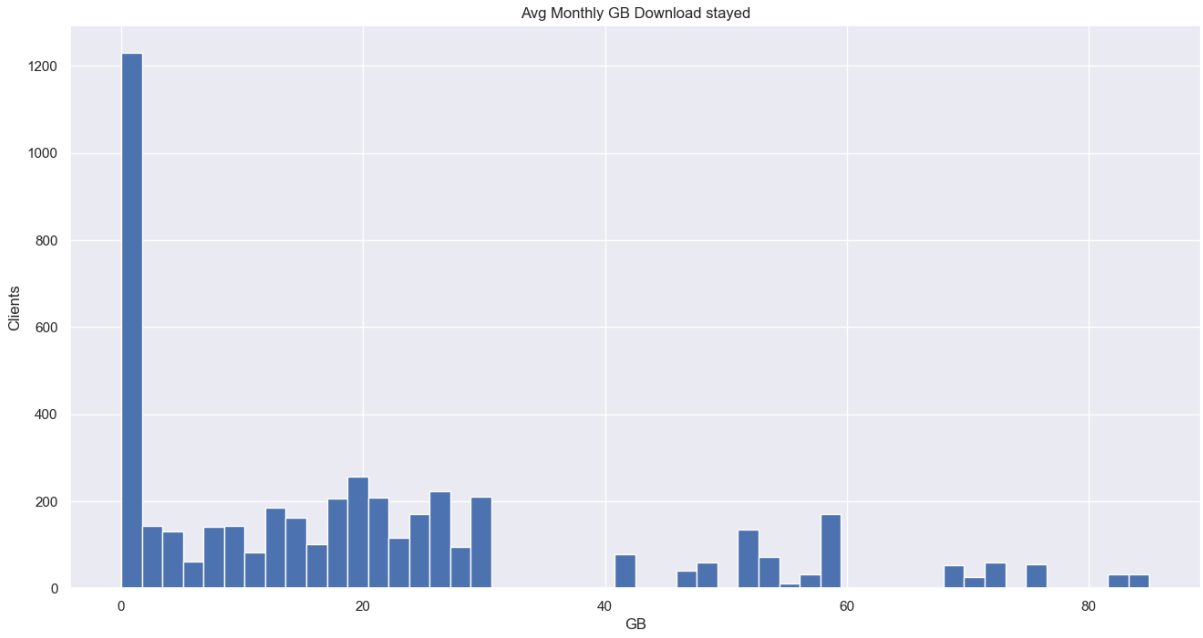

Figura 92: Avg Monthly GB Download Stayed

Els clients que es queden descarreguen menys GB que els clients que marxen. De fet, la major part dels que no descarreguen res es troben entre els clients que es queden.

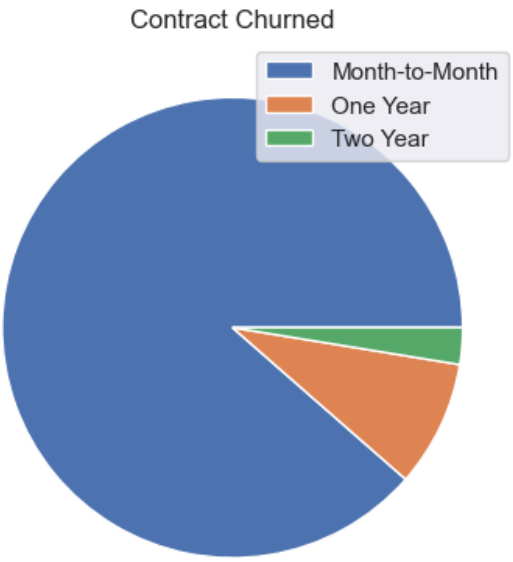

Figura 93: Contract Type Churned

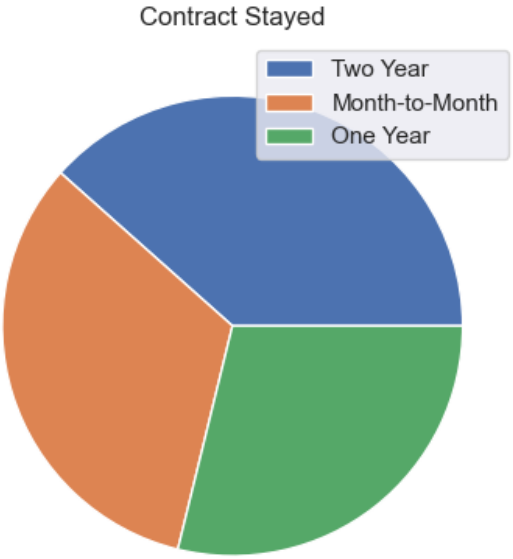

Figura 94: Contract Type Stayed

La majoria de clients que deixen la empresa paguen mes a mes.

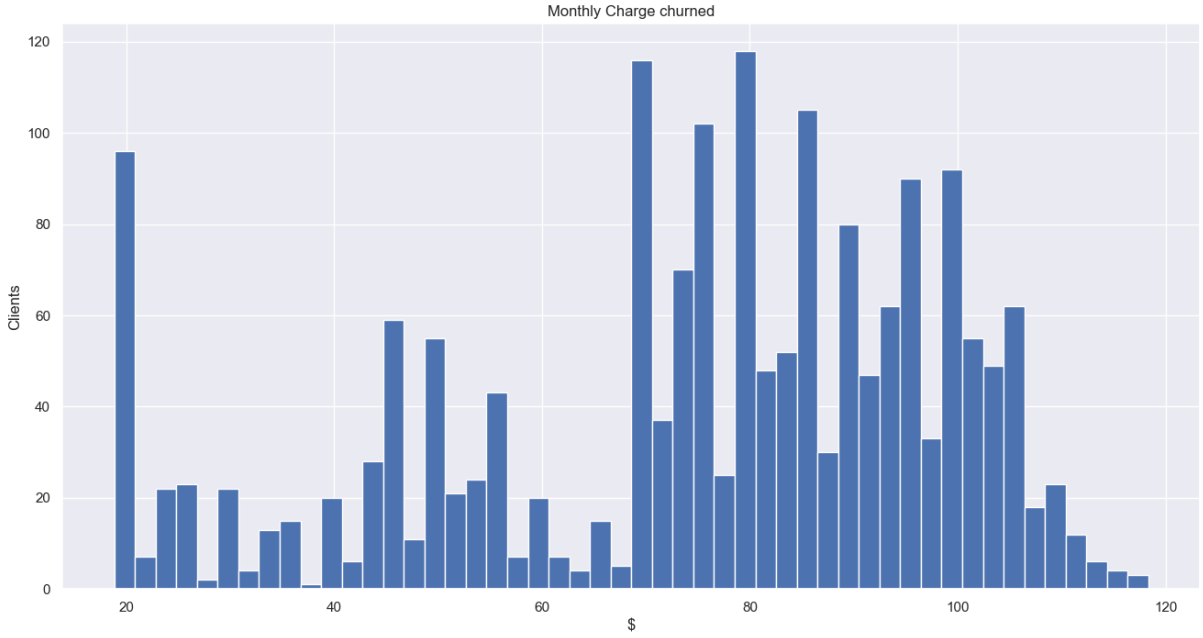

Figura 95: Monthly Charge Churned

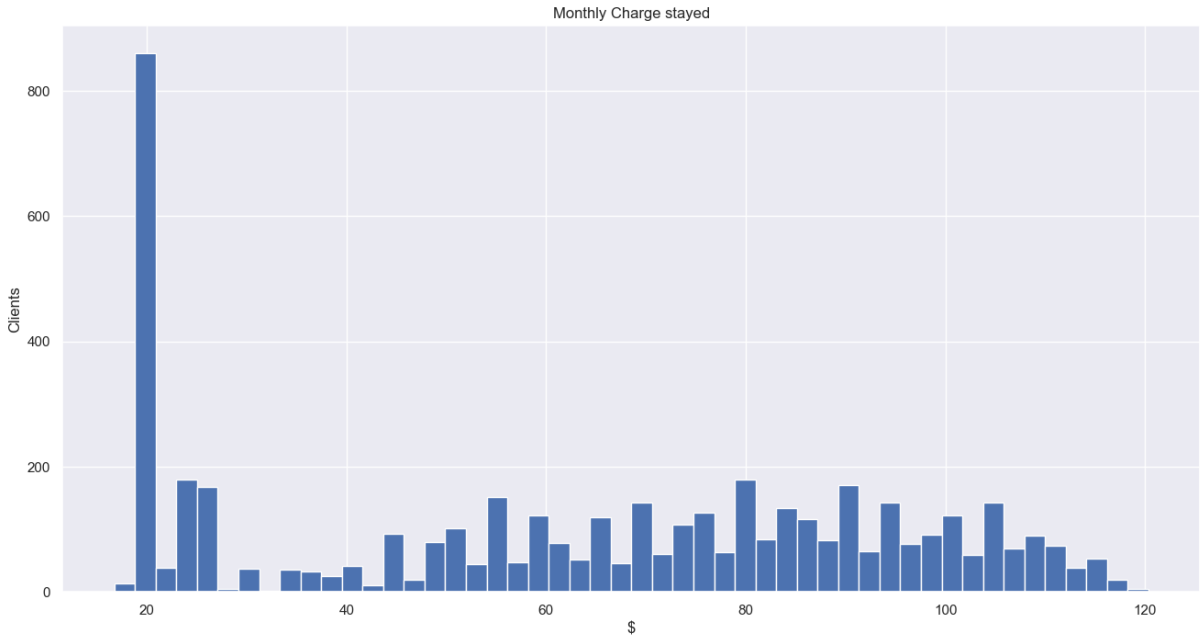

Figura 96: Monthly Charge Stayed

La majoria dels clients que es queden gasten menys que els clients que marxen. Sembla que els clients que marxen tenien tarifes més elevades. La distribució dels que marxen està desplaçada cap a la dreta.

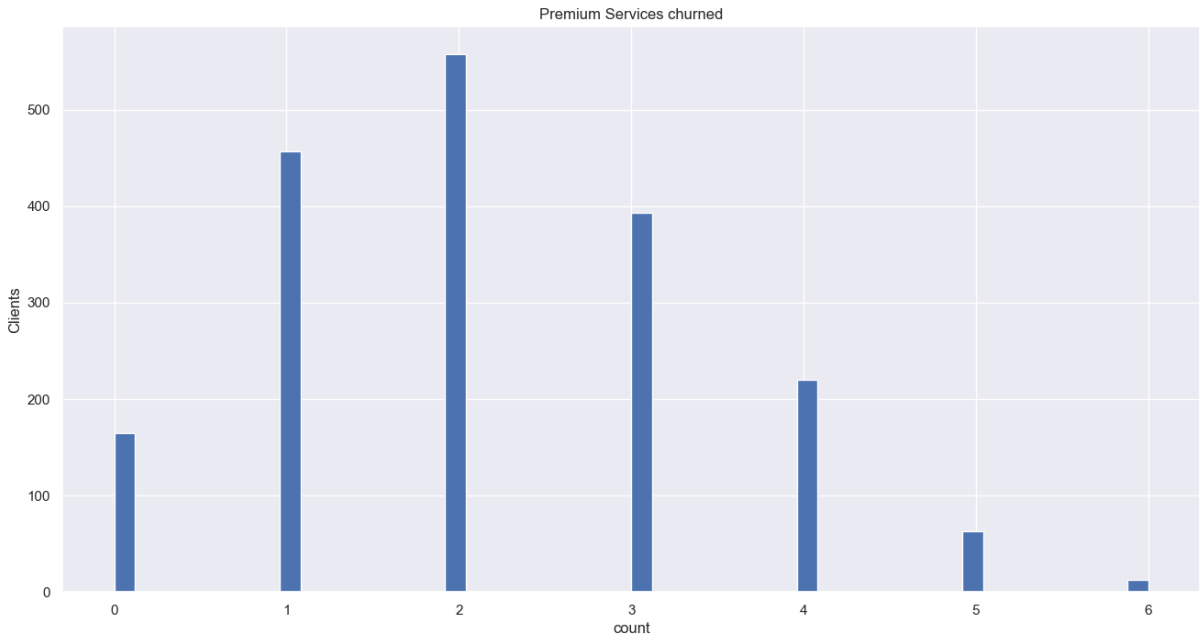

Figura 97: Premium Services Churned

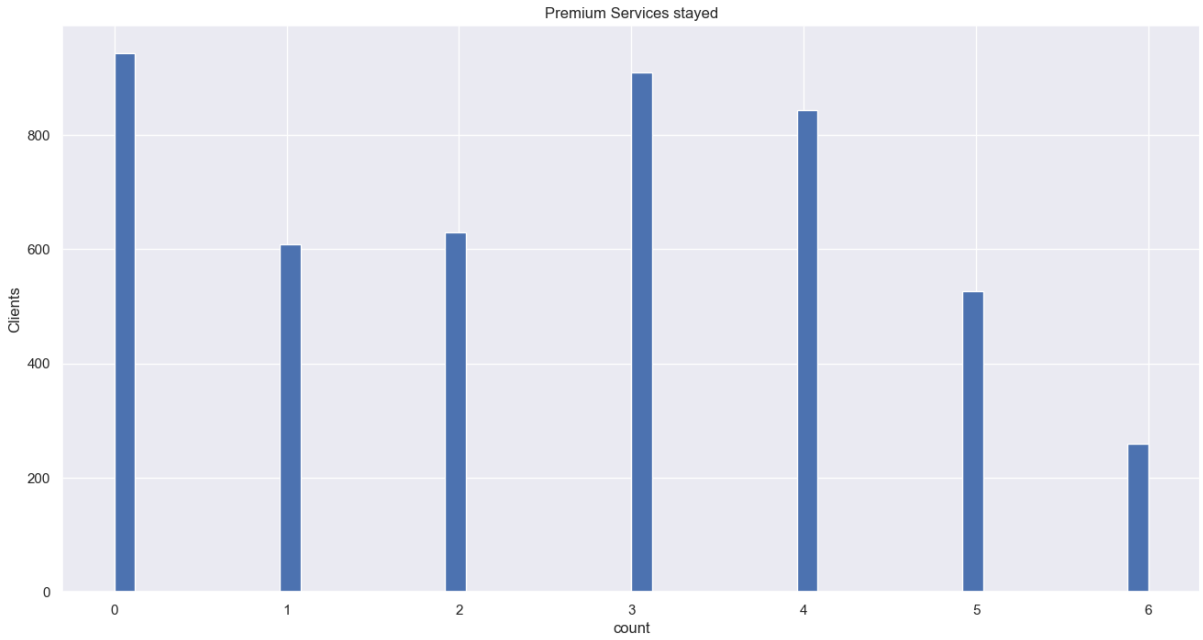

Figura 98: Premium Services Stayed

Les distribucions són diferents. En els extrems es nota més ja que pels clients stayed 0 és on n'hi ha més i proporcionalment n'hi ha més en els valors més grans. En canvi, pels clients que marxen el centre hi ha més clients.

Conclusions

Els clients que marxen paguen més, són propensos a tenir més edat, tenen molts més càrrecs per passar-se del límit de dades, tendeixen a tenir fibra òptica i a pagar mes a mes.
En canvi el clients que es queden porten més temps, són més propensos a estar casats, tenen més serveis premium contractats i són propensos a referir un conegut i tenir dependents.

Per prevenir s'ha de vigilar amb els clients que paguen mes a mes i no porten tant temps dins la companyia. Si apart són clients amb les càrregues mensuals altes, llavors potser s'hauria de fer una oferta millor que inclogui més dades o que el preu d'utilitzar dades de més sigui menor i aprofitar perquè paguin cada any per exemple, ja que si accepten això possiblement és perquè no tenen problema en quedar-se en la companyia. També pot ser que el servei de fibra òptica no sigui el millor del sector i per tant s'hagi de millorar.
Els clients tipus família sembla que tendeixen més a quedar-se i per tant potser a la companyia li convé enfocar-se a aquest tipus de clients.

5 Preparació de dades: Classifier

5.1 Estudi de balanceig respecte la variable objectiu

Si mirem un barplot de Customer Status podem veure que hi ha més clients que es queden que no pas clients que marxen. Si ajuntéssim els clients que acaben d'unir-se a la companyia, llavors el desbalanceig es faria més gran. A més a més, com hem dit abans els que s'acaben d'unir poden posar soroll a les dades, així que els eliminem.

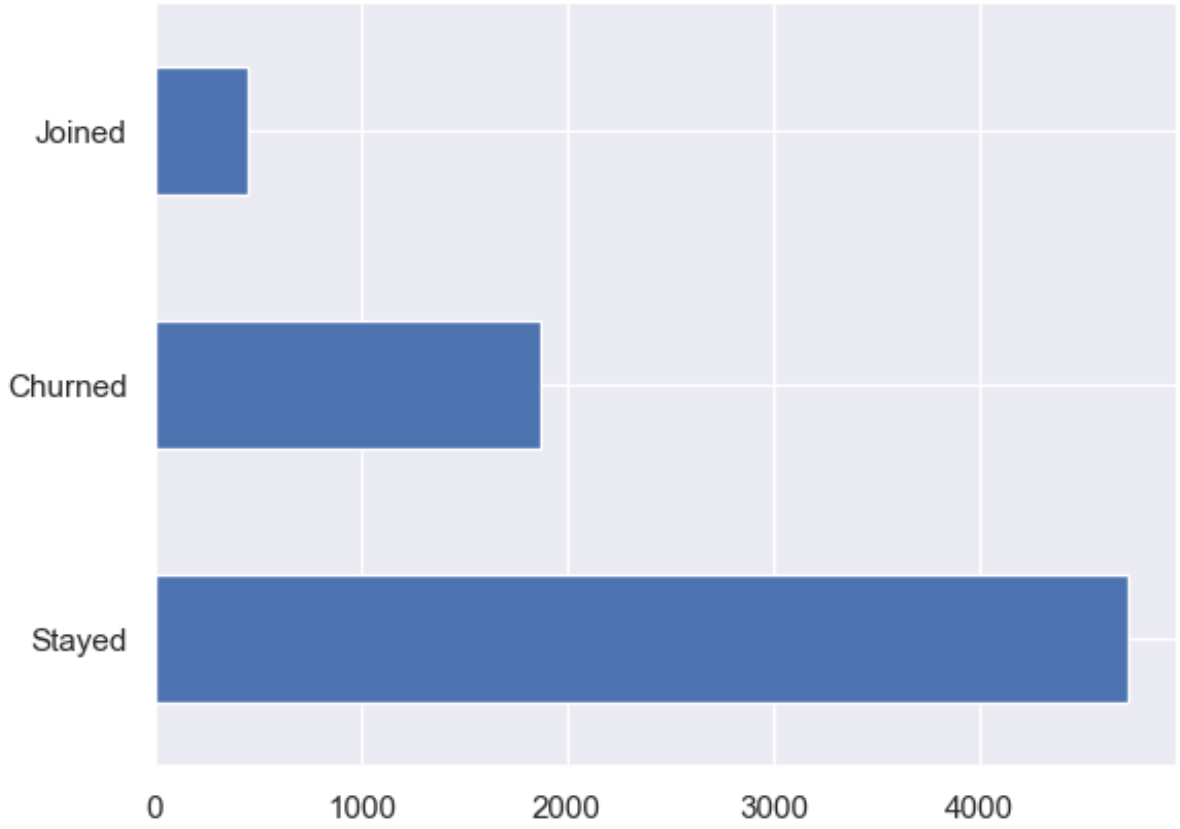

Figura 99: Bar plot Customer Status

Creem una nova variable 'Churn Status' que serà si el client es queda (customer) o marxa (churned) i borrem Customer Status. La nova variable té el següent balanç:

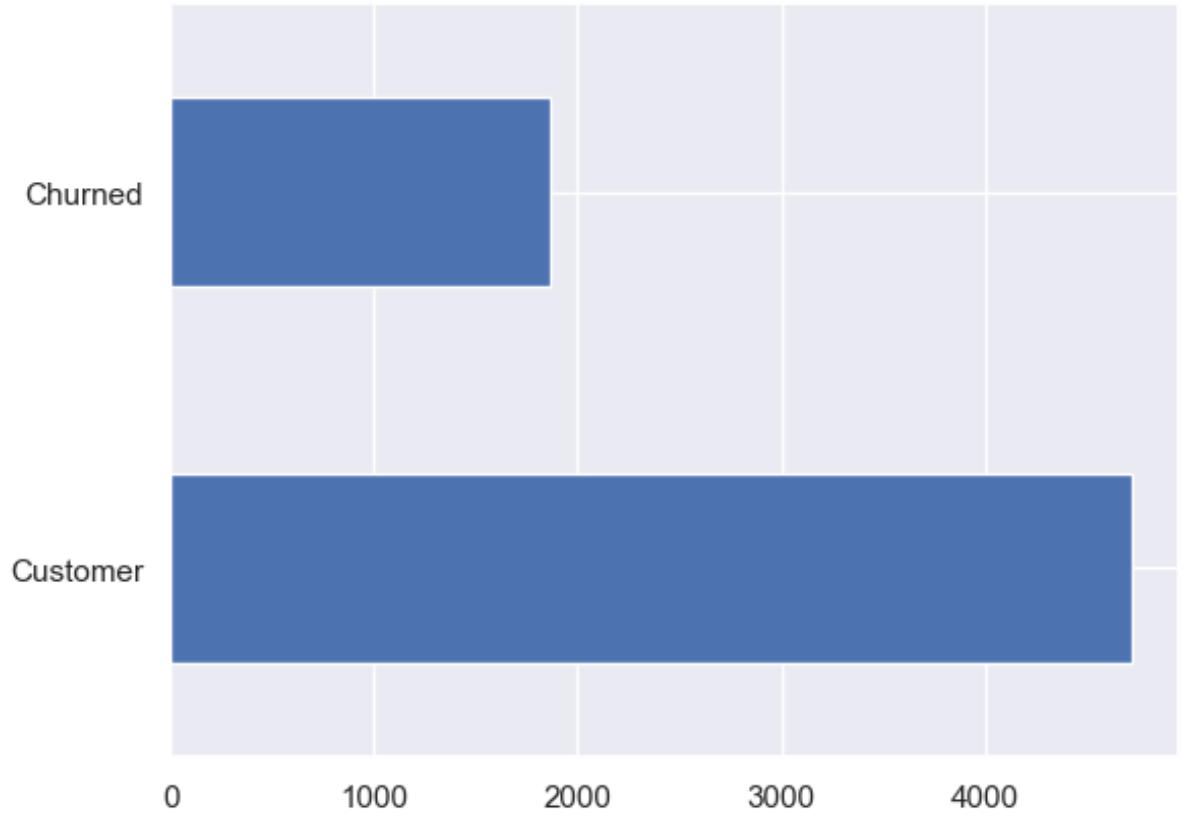

Figura 100: Bar plot Churn Status

La variable continua estant molt desbalancejada i per tant s'haurà d'aplicar algun mètode per mitigar-ho.

5.2 Partició

Volem que el test estigui perfectament balancejat respecte la variable objectiu per poder observar resultats més precisos per saber si el model funciona suficientment bé i sobretot perquè aprengui patrons de les dades i no el desbalanceig d'aquestes.
El particionat el volem 80% per entrenar, 10% per validar i 10% pel test final. Com més dades tinguem en l'entrenament millor, però per poder tenir un test relativament veraç ens hem de reservar una part. Per altra banda volem que el subset de validació tingui les mateixes proporcions que el test i per tant estigui perfectament balancejat. Al agafar un 20% de les dades perfectament balancejades respecte Churn Status (10 validació i 10 de test), la modalitat menys representada (Churned) tindrà encara menys proporció de representació en les dades de train.
És a dir que la resta de dades estaran inclús més desbalancejades.
Per aconseguir aquests resultats hem creat una funció pseudo-aletòria per fer l'split. Aquesta funció ens permet especificar la variable respecte la qual balancejar el test i quina proporció de les dades totals ha d'anar al test i llavors fa un shuffle per no agafar les dades ordenades i després les separa.

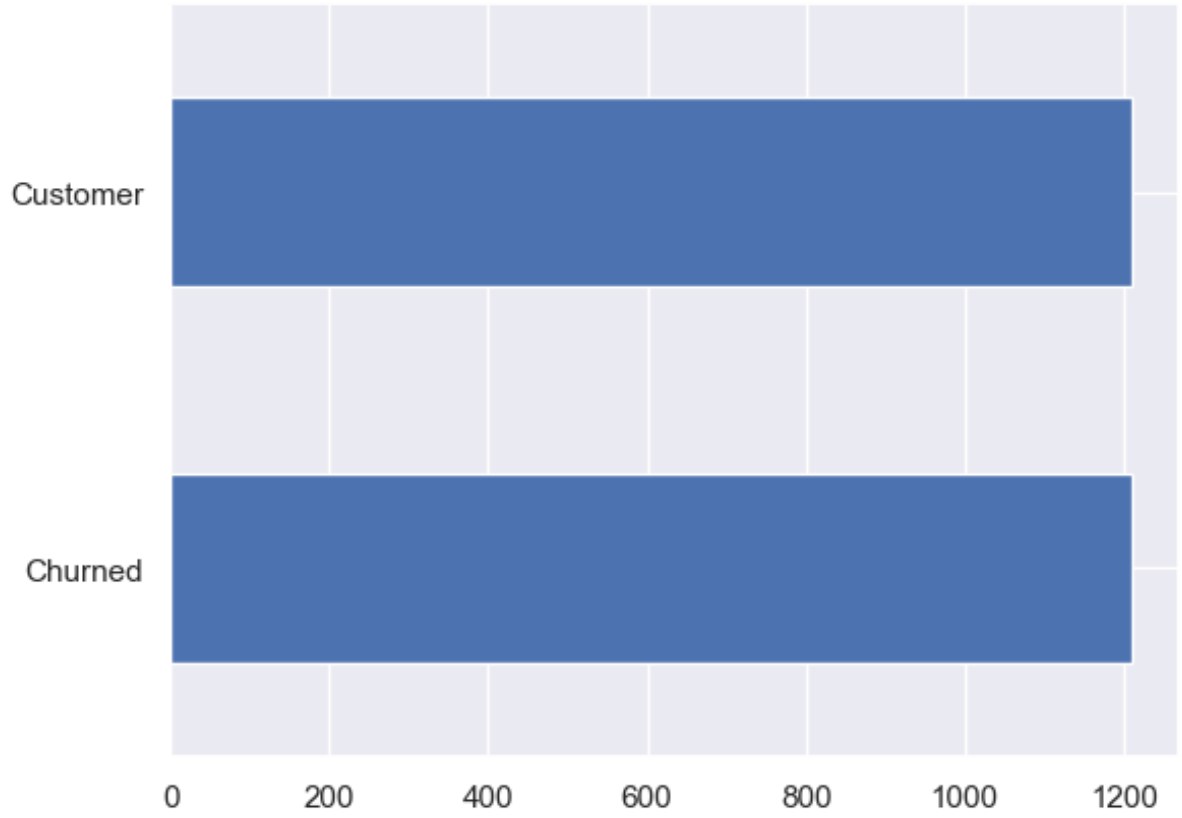

Figura 101: Balanç Churn Status Test+Validation

5.3 Estratègia mitigació del desbalanceig en train-val

Per mitigar aquest desbalanceig ens plantegem tres possibles solucions, les quals es podrien inclús fusionar: Undersampling, oversampling, cross-validation.

Oversampling

Oversampling es basa en crear noves files de les dades que tenim per balancejar una variable. Al utilitzar algun mètode d'oversampling estas creant individus teòrics que en la realitat no existeixen o estas repetint individus que ja tens en les dades i que per tant els estas assignant un valor més elevat. És per això que no utilitzarem tècniques d'oversampling.

Undersampling

Les tècniques d'undersampling es basen en reduir les files per balancejar una variable, en el nostre cas la variable objectiu. Com em dit prèviament, com més dades utilitzem per entrenar, teòricament és millor, així que eliminar moltes files no convenç en aquest cas.

Cross-Validation

Amb cross-valiation no tens un split fix de train validation, per exemple en un 4-fold cv tens 4 splits de les dades de train i llavors proves d'entrenar amb totes les combinacions deixant un split per validació. Al fer això representa que aprens de les dades en general i no només de les d'entrenament, és a dir que redueixes la possibilitat d'overfit.

Volem utilitzar CV, llavors tenim 2 opcions, fem un CV personalitzat on el split de validació estigui balancejat respecte Churn Status ja que les proporcions del test i validation han de ser les mateixes. L'altre

opció és fer undersampling amb les dades que utilitzarem pel CV. No podem fer oversampling amb CV perquè llavors si hi ha dades sintètiques en la part de validació estaríem avaluant un model amb clients que potser no poden existir (per exemple que estigui sota dues ofertes alhora). Utilitzarem 2 mètodes d'undersampling. Tomek links i random undersampling. El primer ens permet eliminar clients stayed que s'assemblen molt a churned i mantenir les diferències. El problema d'això és que encara no ens cobreix tot el desbalanceig ja que no queden més links que eliminar. Així que després fem random, que pots eliminar tot el que vulguis. Al final ens queda el train-val, el validation i el test perfectament balancejat.

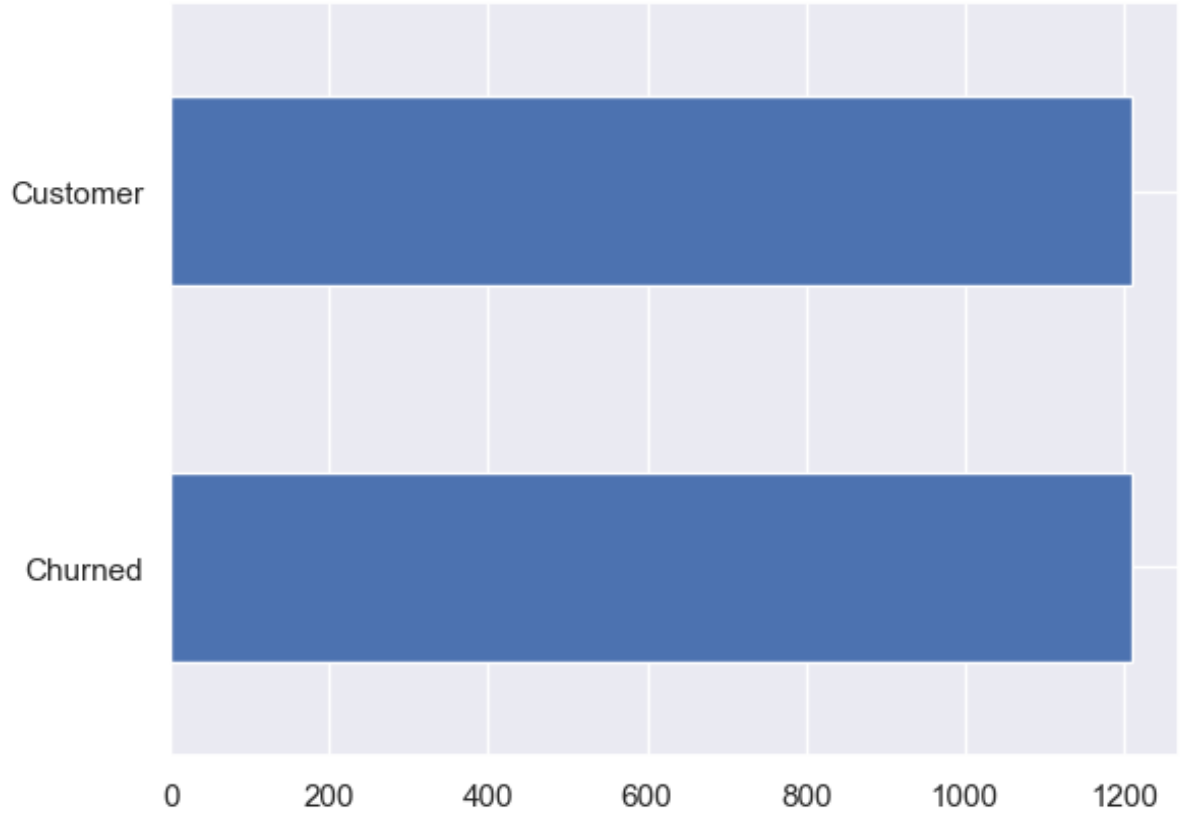

Figura 102: Balanç Churn Status Train després d'undersampling

També hem de tenir en compte que estem fent un subset de validació i cross-validation alhora. Això és perquè utilitzarem el cross-validation per escollir els hiperparàmetres dels models i després la partició fixa de validació per escollir entre els diferents models.

Al final el procediment que hem seguit és:
fer un split del 20% que després seprararem entre el test i la validació. Fer undersampling del 80% restant. Fer split del 20% anterior entre test i validation (50/50). Entrenar els models amb cross-validation amb 4 folds mitjançant GridSearchCV de sklearn per escollir els paràmetres. Provar la configuració escollida de cada model amb sobre les dades de validació fixa. Utilitzar aquest últim resultat per escollir el model definitiu. Provar aquest model en el test.

5.4 Normalització

Les transformacions es fan només amb el coneixement de les dades de train. És a dir que ajustem les transformacions amb les dades de train i després transformem train, test i val amb la mateixa transformació.

Estratègia general

La estratègia general és:

- Normalitzar mitjançant BoxCox i escalar mitjançant minmax scaler pel KNN
- Deixar tal qual pels arbres
- Escalar mitjançant minmax scaler per SVM

I fer OneHot encoding per totes les variables categòriques.

Aquí podem veure que el boxcox no ha pogut acabar de donar forma de campana de Gauss a les variables.

Figura 103: Formes distribucions variables KNN

6 Definició de models

En aquesta secció s'entrenarà 6 models diferents. Un KNN, un decision tree, un random forest , un XGBOOST i dos SVM, un lineal i un amb kernel rbf. Amb el cross-validation i mètriques que ara definirem es decidirà els hiperparàmetres dels models. Volem assegurar-nos que escollim el màxim dels clients que marxen, per tant tindrem molt en compte el Recall dels churned. Per altra part, no volem simplement dir que tots són churned, així que podríem utilitzar la mitjana harmònica de la precisió i el recall per valorar els models. Finalment, com les dades estan totalment balancejades, aquí podrem utilitzar accuracy per valorar el model. Després també farem matrius de confusió amb els paràmetres escollits amb el CV provats en les dades de validació fixa.
Bàsicament, el cross-validation farem que es guii per Recall ja que volem minimitzar el nombre de churned que no classifiquem com a tal. Per altra banda, després amb la segona validació mirarem totes les mètriques i la matriu de confusió tenint en compte que volem minimitzar el nombre de churned que classifiquem malament. Els TP són els churned que predeim com a churned i els FN són els churned que predeim com a stayed. Això ens serà de molta ajuda per analitzar les matrius de confusió.

6.1 Mètriques

$$ACCURACY = \frac{TP + TN}{P + N}$$

$$PRECISION = \frac{TP}{TP + FP}$$

$$RECALL = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{PRECISION \times RECALL}{PRECISION + RECALL}$$

6.2 KNN

Motivació del model

KNN és molt interpretable si les dades no estan transformades, en el nostre cas hem hagut de normalitzar i escalar, de forma que perdem la interpretabilitat de KNN. Per altra banda, té relativament pocs hiperparàmetres, principalment la k, no obstant també es pot canviar la definició de distància i el pes dels individus. El principal motiu pel qual utilitzem KNN és per marcar una baseline per comparar la resta de models.

Hiperparàmetres

K rang(1, 50) Volem valors ni molt petits ni molt grans, limitem a 50 ja que 50 de 2420 és poc. Inclús potser hauríem de provar valors més alts. Ponderació per distància (si/no) Ponderar per distància pot donar millors resultats però és més fàcil fer overfit. P: Potència de Minkowski (p=2 és el mateix que la distància euclídia); només provarem 1 i 2.

Resultats experiment

Hem executat GridSearchCV amb les especificacions anomenades prèviament hi ha donat uns resultats de 0.8768414093064936 de recall. Els hiperparàmetres escollits han estat: p=2, k=46 i ponderat per distància. Tenint en compte el número de files de les dades d'entrenament 46 tampoc és tant. Ponderar per distància típicament treu millors resultats, però el cost és que és més fàcil fer overfit, com hem utilitzat CV confiem en que això no ha passat.

Resultats finals

Al provar el model amb les dades de validació fixa hem obtingut un recall de 0.87 i una accuracy del 80%.

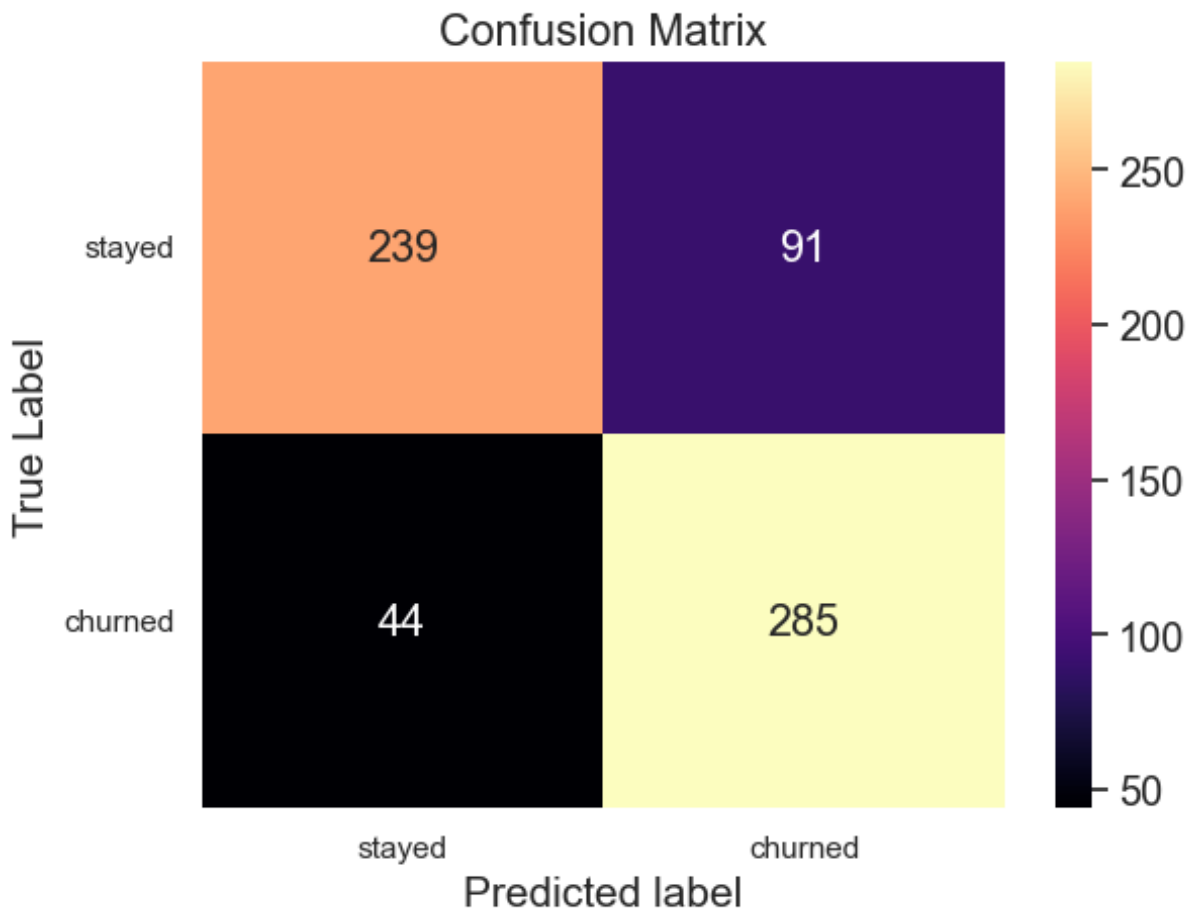

Figura 104: Matriu de Confusió KNN

Tenim 285 TP i 44 FN.

6.3 Arbre de decisió

Motivació del model

Un arbre de decisió té moltes avantatges. Com no utilitza distàncies no hem de fer càlculs de distàncies que poden ser problemàtics i ens traiem de sobra la 'curse of dimensionality' i ja no hem de escalar ni normalitzar i ens permet introduir variables categòriques. Per altra part, té hiperparàmetres, però aquests serveixen per evitar overfitting. A més a més, és molt interpretable ja que no és res més que una sèrie de 'if'. La part negativa és que és sensible a dades desbalancejades i a datasets petits, cosa que en aquesta ocasió no tenim, el dataset tampoc és molt gran però disposem de 2400 files pel CV.

Hiperparàmetres

Profunditat màxima rang(1,50) amb profunditats màximes grans l'arbre fa overfit, però si la profunditat és molt petita hi ha underfit. Mínim d'individus per fulla rang(1, 30) amb mínim per fulla petit tendeix a fer overfit. Criteri per fer els splits: gini i entropy (els dos són semblants en la majoria de ocasions, però els podem provar).

Resultats experiment

El resultat del recall del CV és 0.8818164929075689. I els paràmetres que l'han aconseguit són: profunditat màxima 1, mínim de samples per fulla = 1 i criteri gini. Gini és el criteri més utilitzat i que millors resultats m'ha donat així que no és inesperat. La questió és que amb profunditat màxima 1, llavors només és com un if basat en una sola variable.

Resultats finals

En la validació fixa hem obtingut un recall del 0.9 i una accuracy de 0.79

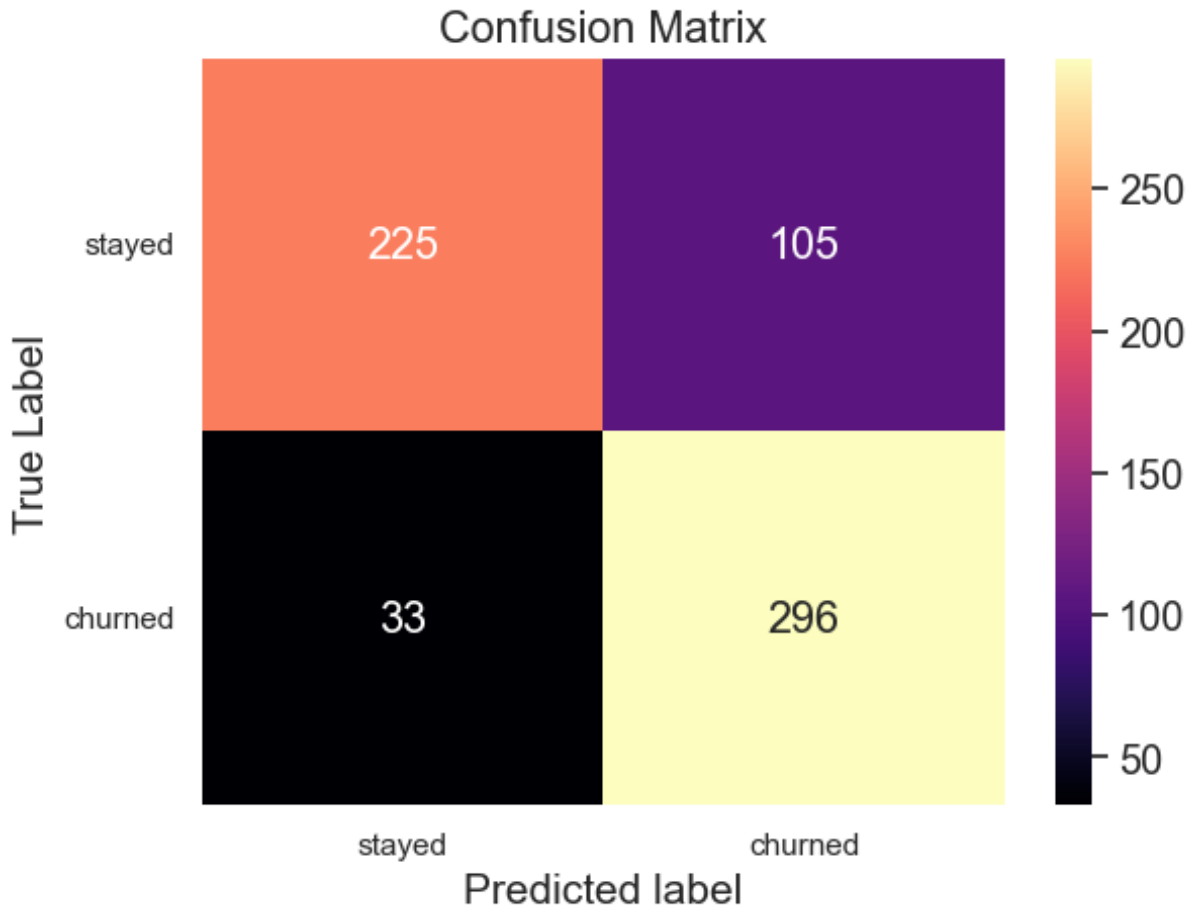

Figura 105: Matriu de Confusió Decision Tree

Ha augmentat els TP i disminuït els FN però també ha augmentat els FP i disminuït els TN. L'arbre que ha obtingut aquests resultats és un sol if que podem analitzar.

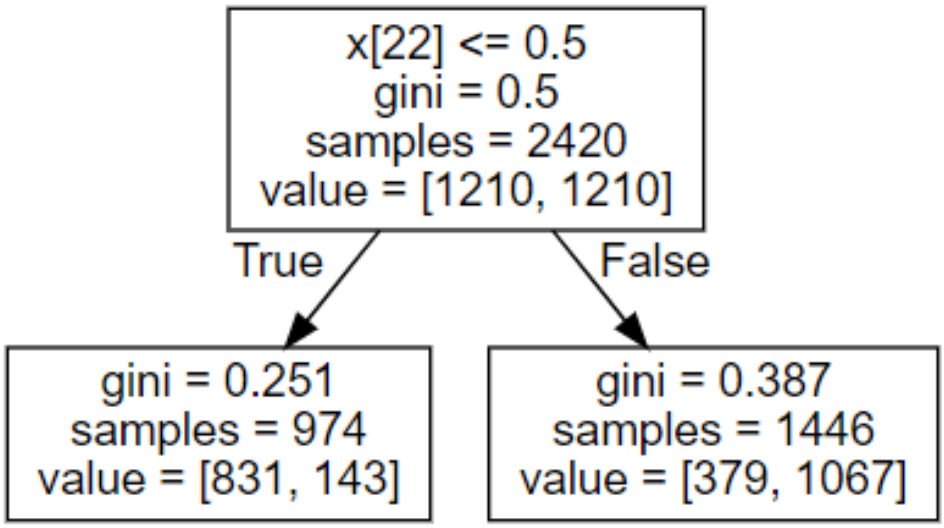

Figura 106: Decision Tree

Parteix basat en si el client paga mes a mes o no. Els clients que paguen mes a mes són els que marca

com a churned. Això té sentit perquè si saps que si pagues cada mes temps és perquè saps que no marxaràs. No obstant, malgrat la seva interpretabilitat, em sembla un model massa senzill, ja que llavors cada mes hauries de trucar a tots els clients que paguen mensualment per evitar que se'n vagin.

6.4 Random forest

Motivació del model

En el nostre cas, el dataset potser és molt reduït per treure bons resultats ja que hem aplicat molta reducció amb undersampling. Fent un Random Forest representa que agafem els aspectes positius dels arbres de decisió i afegim que al fer bagging, llavors potser treu bons rendiments malgrat el tamany reduït de les dades i com els models són independents es fan alhora. Anteriorment hem vist que amb un arbre molt senzill ja obtenim resultats acceptables, així que una combinació d'aquests potser funciona millor.

Hiperparàmetres

Profunditat màxima Mínim d'individus per fulla Nombre d'arbres creats: si només és un és el mateix que un decision tree, si n'hi ha molts potser uns arbres que treuen pitjors resultats tapen als arbres més útils.

Resultats experiment

El CV ens ha donat un recal de 0.91. Els hyperparàmetres escollits són: profunditat màxima 1, mínim samples per fulla de 33 i 5 arbres. Torna a passar que els arbres són un sol if. Però aquest cop s'ha modificat el mínim per fulla. Amb 5 arbres el model encara és força interpretable, tenint en compte que són de profunditat 1.

Resultats finals

Els Random Forest poden treure bons resultats en una partició i no tan bons en un altra fàcilment. Aquí ha passat, en la validació fixa ha tret un recall de 0.78 i una accuracy de 0.8.

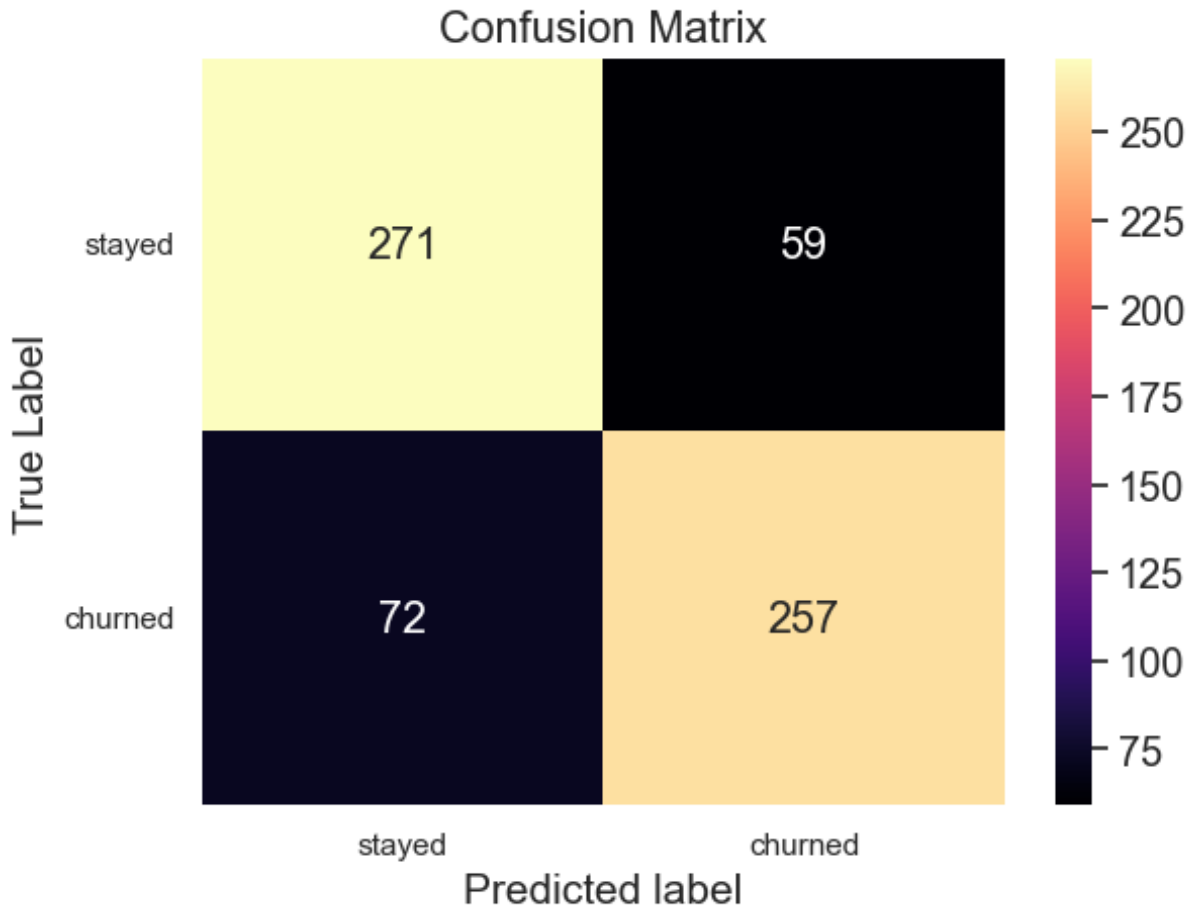

Figura 107: Matriu de Confusió Random Forest

No ens serveix pel nostre objectiu. Inclús prediu millor els stayed que els churned.

6.5 XGBOOST

Motivació del model

Partim dels factors positius del arbres de decisió i afegim que ara fem un arbre a partir de l'arbre anterior basant-nos en els seus errors moltes vegades, de manera que al final aconseguim un arbre millorat.

Hiperparàmetres

Profunditat màxima: range(1,31, 5) Mínim d'individus per fulla: range(15,36, 3) Nombre d'arbres creats: range(1,25) Learning rate: quan volem que canvi un arbre respecte el següent arbre. [0.0001, 0.001, 0.01, 0.1, 0.5]

Resultats experiment

Hem obtingut un recall de 0.8826415754158198 amb els següents hyperparàmetres:
Learning Rate = 0.5, profunditat màxima 1, mínim samples per fulla 15 i 2 arbres. El learning rate és el màxim, és a dir que mig destrueix el primer arbre. Només fa dos arbres i són de profunditat 1 així que el model és molt interpretable. Però potser no és un model suficient específic.

Resultats finals

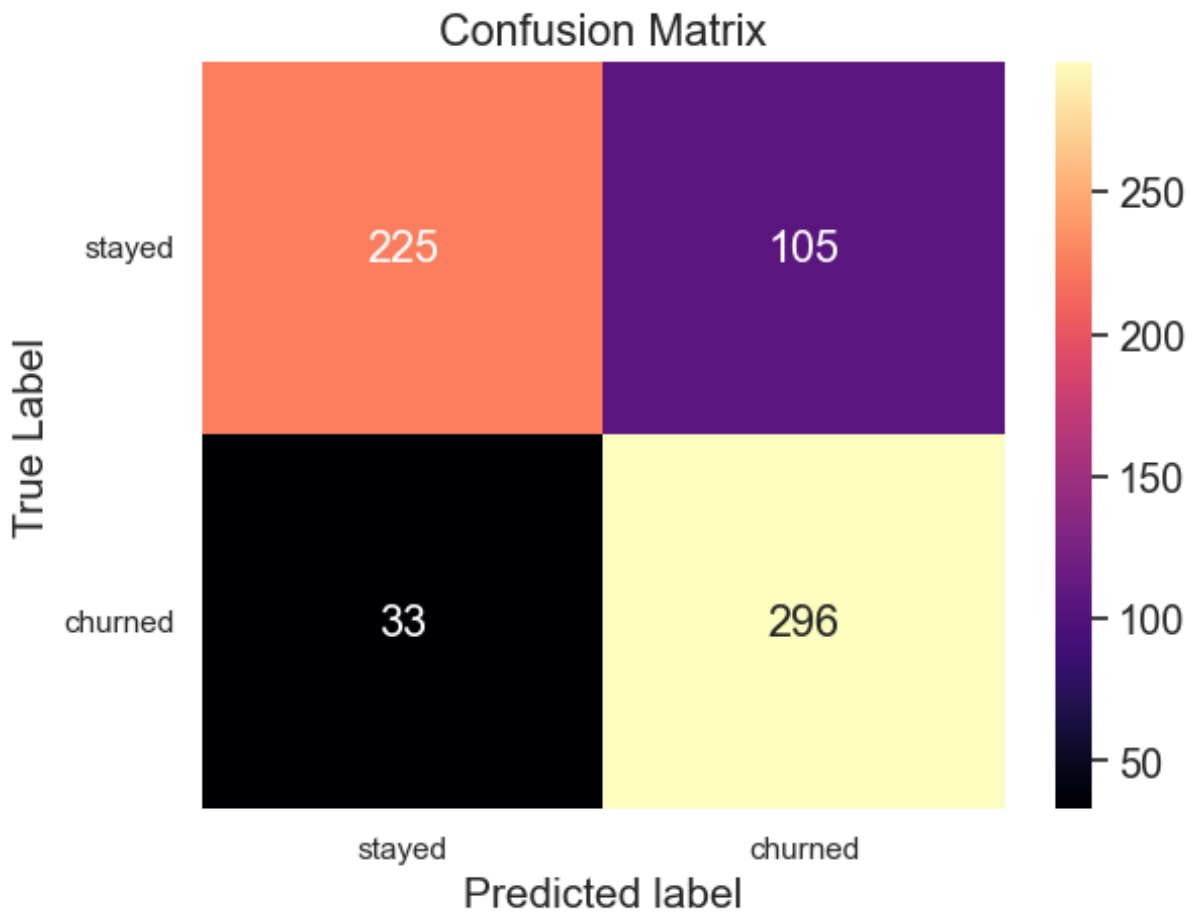

Figura 108: Matriu de Confusió XGBoost

Treu els mateixos resultats que el decision tree, i tenint en compte els paràmetres del xgboost potser fa el mateix split en el 2n arbre i com el learning rate és 0.5 es carrega massa el primer de forma que dona els mateixos resultats que el decision tree.

6.6 SVM

Motivació del model

Les parts bones dels SVM és que són resistents als outliers així que no ens hem de preocupar de clients anòmals, que són deterministes de forma que sempre dona el mateix resultat i no té cap assumpció gaussiana, de manera que no hem de estandarditzar les dades. No obstant, escalem per no donar més pes a una variable, ja que sinó la configuració del núvol de punts en l'espai seria diferent.

Hiperparàmetres

C: serveix per decidir si prioritzar els marges (C baixes) o l'error (C altes). [10000.0, 1000.0, 100.0, 10.0, 1.0, 0.1, 0.01, 0.001, 0.0001]

Resultats experiment

El cross-validation treu un recall de 0.8818055646624265 amb C=0.001.

Resultats finals

En la 2a validació el recall és de 0.91 i la accuracy de 0.8.

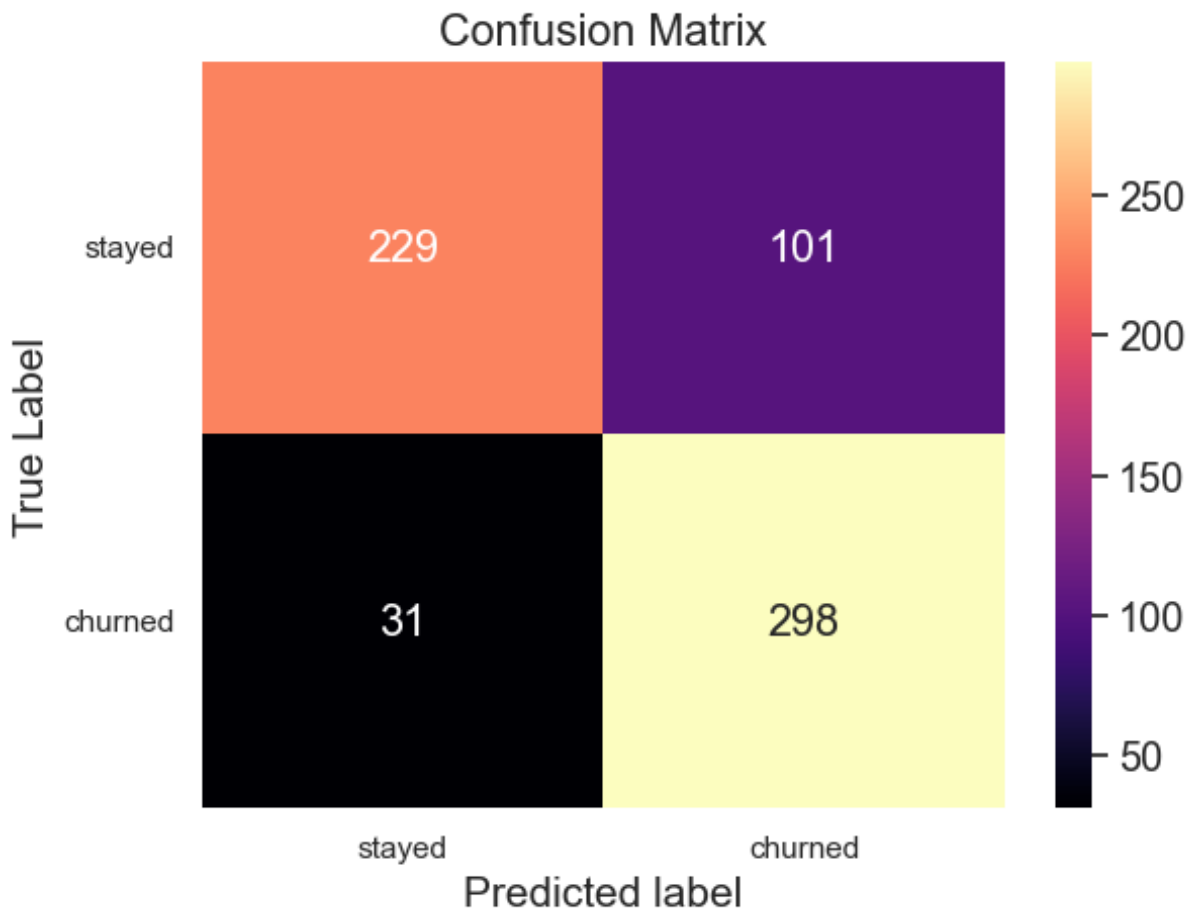

Figura 109: Matriu de Confusió SVM lineal

És el model que més TP ha tret i que té tota la pinta de ser l'escollit.

6.7 SVM rbf

Motivació del model

Aquest model no només té les parts positives del SVM lineal, sinó que apart ja no és lineal, així que podem fer separacions no lineals.

Hiperparàmetres

C: [100000, 10000, 1000, 100, 10, 1, 0, 0, 0, 0] Gamma: serveix per ajustar la radial basis function. (gamma baixa underfit / gamma alta overfit) provarem: [100000, 10000, 1000, 100, 10, 1, 0, 0, 0, 0]

Resultats experiment

El millor resultat del cross-validation és el recall de 0.8768496054903504 aconseguit amb C=100 i gamma = 0.0001. No és mala combinació en el sentit que no són valors que portin a overfit o underfit extrems, estan bastant equilibrats.

Resultats finals

En la 2a validació ha tret 90 recall i 80 accuracy.

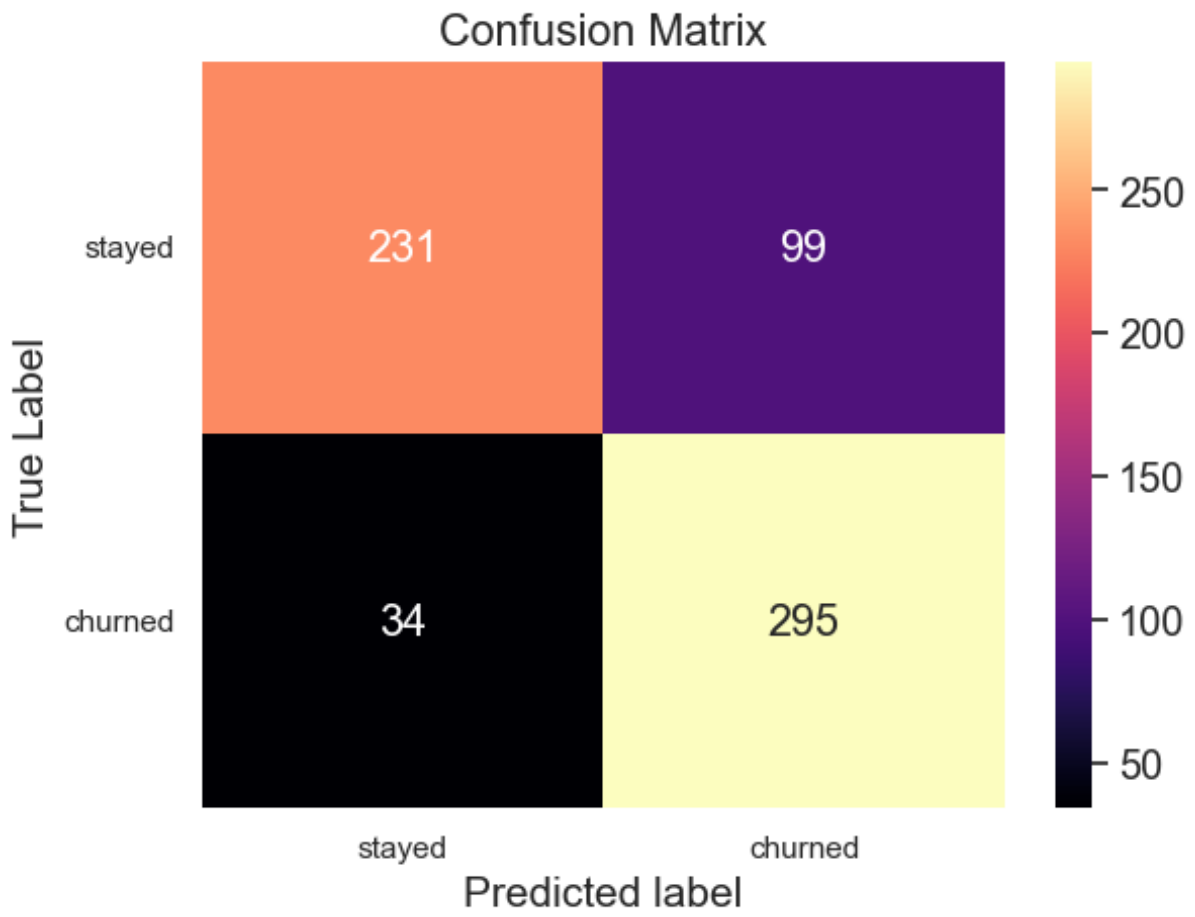

Figura 110: Matriu de Confusió SVM rbf

Podem veure que hi ha 295 TP i només 34 FN. El preu a pagar és que predir stayed no se li dona tan bé.

7 Selecció del model

Decisió

Escollim el SVM lineal amb c=0.001 entrenat amb les dades escalades.

Motius

El motiu principal és perquè és el model que més TP i menys FN fa. L'arbre de decisió podria estar bé, però no crec que preocupar-se per tothom que paga mes a mes sigui una bona idea. Entre SVM lineal i SVM amb kernel rbf estaven bastant empatats, però llavors prefereixo un model més senzill, és a dir el lineal.

Limitacions i capacitats

Com ens he mentalitzat a minimitzar el FN, llavors el model tendeix a predir churned abans que stayed. A més a més el model ha estat entrenat amb dades de Califòrnia i amb uns límits d'edat. Podria ser que en condicions diferents aquest model no funcioni com un desitjaria.

Rendiment

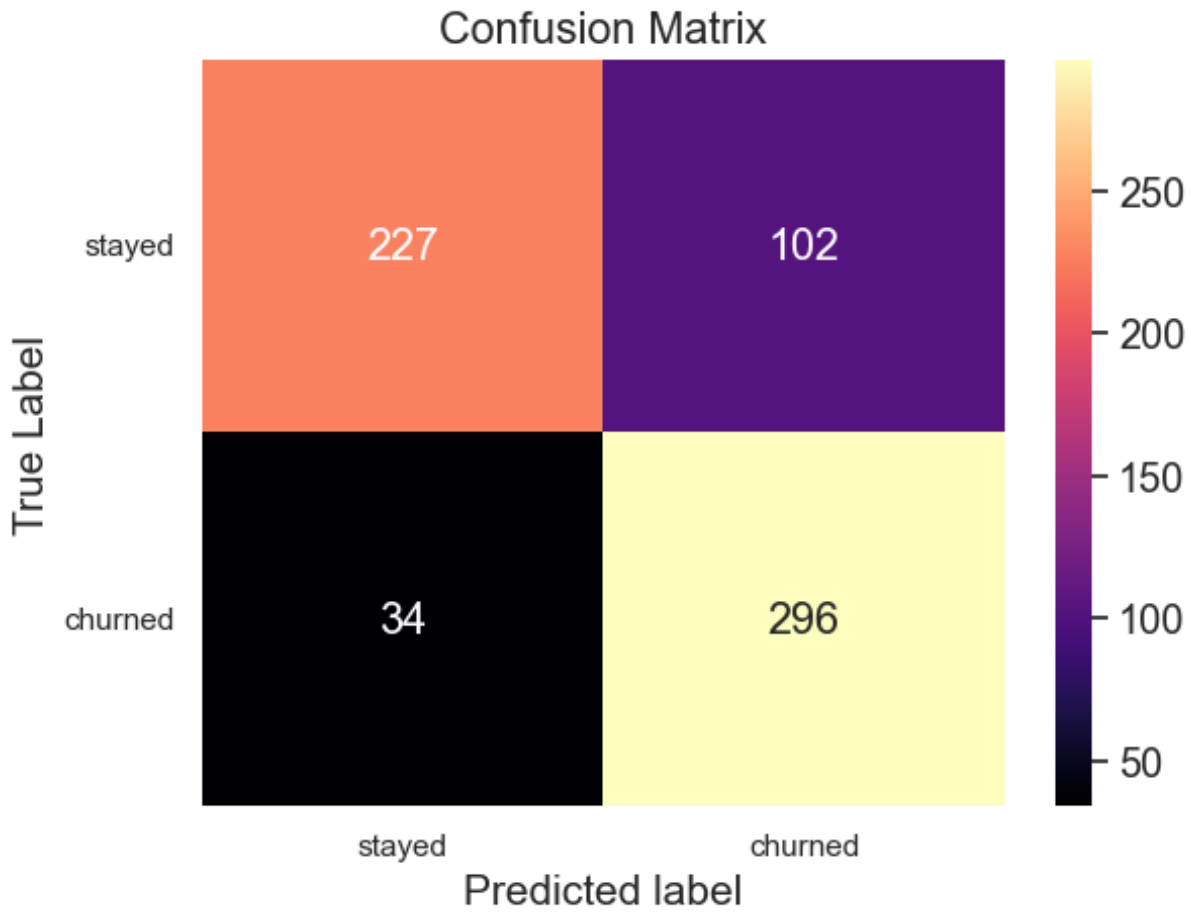

Figura 111: Matriu de Confusió SVM en el test entrenat amb train

Com hem utilitzat doble validació, els resultats de la 2a validació s'assemblen als del test.

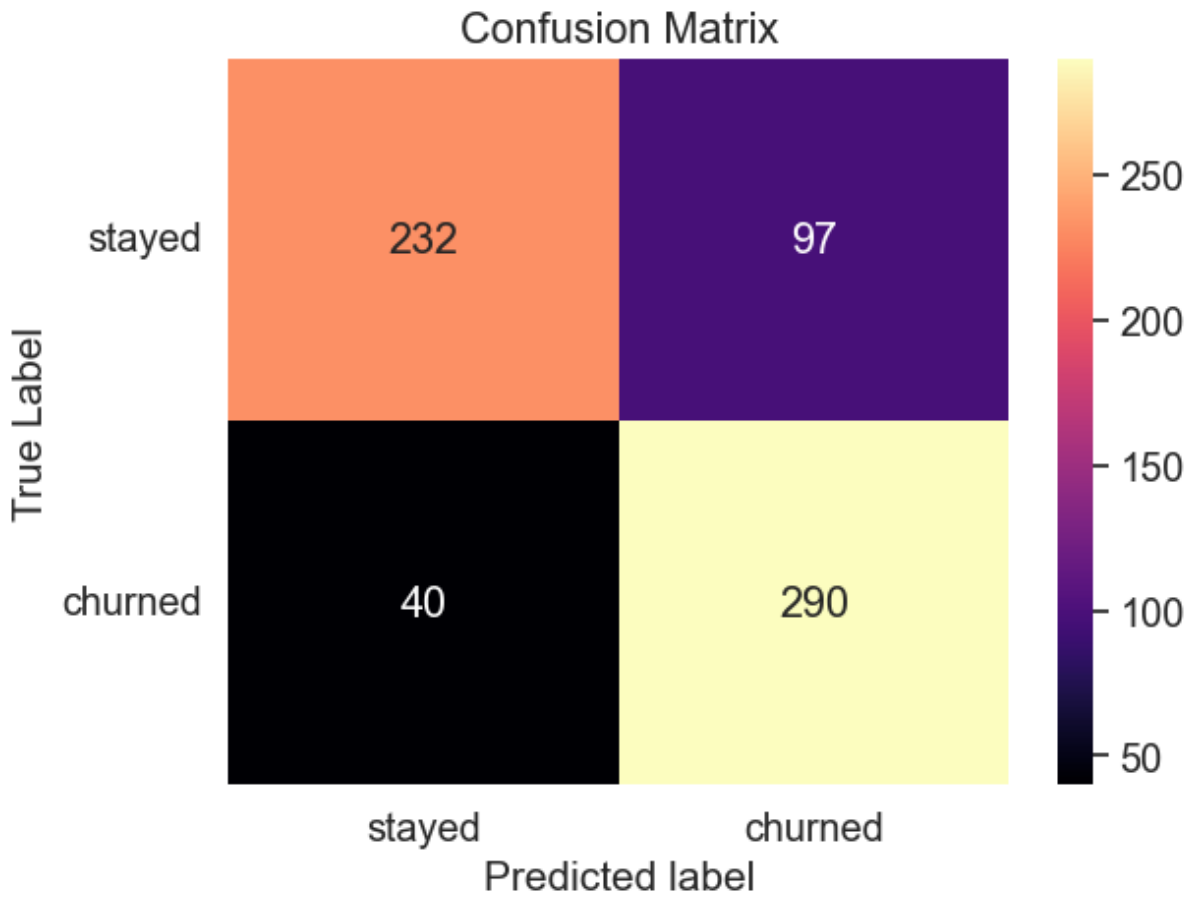

Figura 112: Matriu de Confusió SVM en el test entrenat amb train+val

Al entrenar amb més dades el model no perd rendiment. De fet potser perd TP però prediu més TN. En general sembla que el model no té overfit ni que tampoc dona els millors resultats esperables.

Conclusions

El model prediu els clients que marxen amb una precisió del 0.75. Si ens haguéssim fixat més amb precisió en lloc de recall aquest valor seria més elevat.
Per altra banda, el fet d'haver fet primer split i després undersampling i tenir 2a validació ha fet que les dades d'entrenament minvessin i potser hem perdut capacitat predictiva.

8 Model Card

8.1 Model details

- Desenvolupat per Sergi Guimerà, estudiant de GIA
- Data 2022/2023
- El model és un classificador SVM amb kernel lineal
- S'ha utilitzat la llibreria de sklearn pels algorismes. Els paràmetres escollits han estat C=0.001 realitzada mitjançant cross-validation guiat per recall.
- En cas de dubte contactar a sergi.guimera@estudiantat.upc.edu

8.2 Intended use

- Aquest model s'ha desenvolupat exclusivament per propòsits educacionals.
- Intencionat pels alumnes i docents de FIB-GIA
- El model no serveix per clients que s'acaben d'unir ja que no ha estat entrenat amb aquests i perquè no té sentit trucar a un client que acaba d'unir-se per fer-li una contraoferta.

8.3 Factors

- El model s'ha entrenat amb dades de Califòrnia. Un factor important és el lloc d'on provenen les dades ja que no s'ha provat que funcioni per la resta de llocs.
- No s'ha trobat cap factor que perjudiqui el rendiment del model. S'ha eliminat la variable del gènere ja que mitjançant un test de chi quadrat ha sortit que no era dependent de la variable resposta, això pot portar a que el model funcioni millor per un cert gènere ja que aquest pot anar relacionat amb una variable que si hem deixat. Al treure el gènere ens hem tret el problema de quants gèneres hi ha, ja que aquestes dades assumien que n'hi ha dos.

8.4 Metrics

- La mètrica principal és el recall. També s'ha visualitzat les matrius de confusió.
- No tenim llindars de decisió constant, s'ha escollit el model que ha donat millors resultats

8.5 Ethical considerations

- Hem eliminat la variable gènere que era binària. Les dades d'entrenament no tenien persones de més de 80 anys i no tenien cap variable de raça ni cap manera de comprovar que no estava esbiaixada en aquest sentit.

8.6 Caveats

- En el README està especificat on està cada fitxer i que fa aquest.

9 Clustering

Hem decidit utilitzar un mètode no supervisat per veure patrons sobre les dades en general. Per això ens hem decantat per utilitzar hierarchical clustering ja que ens permet visualitzar un dendograma i així poder decidir el tall 'òptim'. Per altra banda utilitzem DBScan ja que el model que hem escollit en l'apartat supervisat era resistent als outliers de la mateixa forma que ho és aquest mètode de clustering. Per poder visualitzar els clusters hem decidit utilitzar MDS ja que ens permet reduir la dimensionalitat a 2 eixos, que és el que necessitem per fer plots.
Pels dos mètodes de clustering utilitzarem totes les files del dataframe, però només les variables que hem utilitzat per l'apartat supervisat, ja que hem eliminat les variables que eren dependents unes de les altres. També trèiem Churn Status. A aquestes dades hi fem minmax scaling per deixar-les amb els mateixos rangs.

9.1 DBSCAN

Per utilitzar DBSCAN podem escollir quina distància utilitzar. La distància de gower ens permet utilitzar variables categòriques sense la necessitat de fer onehot encoding així que escollim aquesta. Aquesta distància ja escala les variables al moment de calcular les distàncies, en el nostre cas no és necessari ja que ja tenim les variables escalades. També hem de decidir els paràmetres pel clústering. DBSCAN té 2 paràmetres, un pel mínim d'observacions que hi ha d'haver en un cluster i l'altre la distància màxima que pot haver entre 2 individus per considerar que són veïns. Com la distància està entre 0 i 1, un bon llindar per la distància màxima pot ser 0.2 aproximadament. Ara per quants clients ha d'haver per fer un clúster, com tenim 7000+ files mínim els clústers haurien de ser de 50. Juguem una mica amb el llindar de la distància i amb 0.1 obtenim 2 clusters i 3622 outliers.

Interpretació resultats

Primer de tot necessitem fer MDS. Per MDS també utilitzarem la distància de gower.

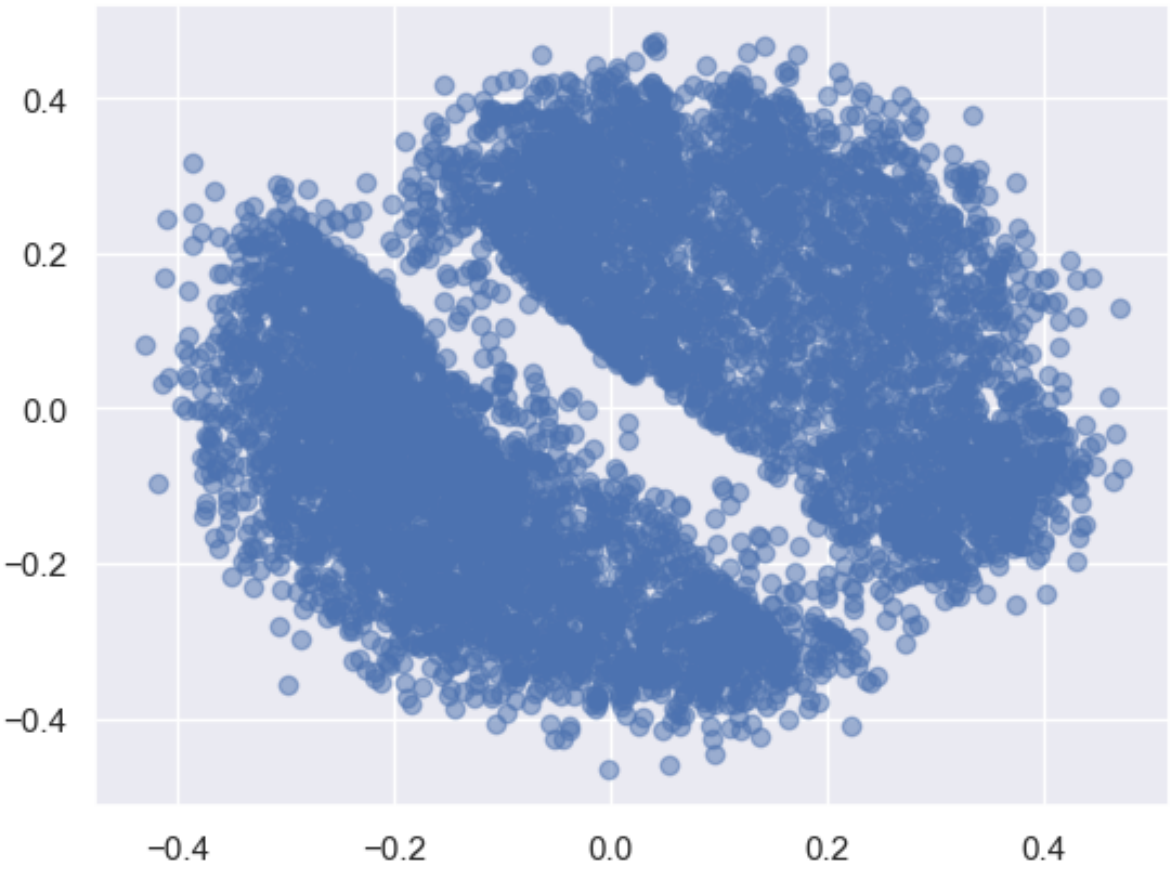

Figura 113: MDS

Podem veure 2 clusters. Estaria interessant que coincidissin amb els clusters del mètodes de clustering.

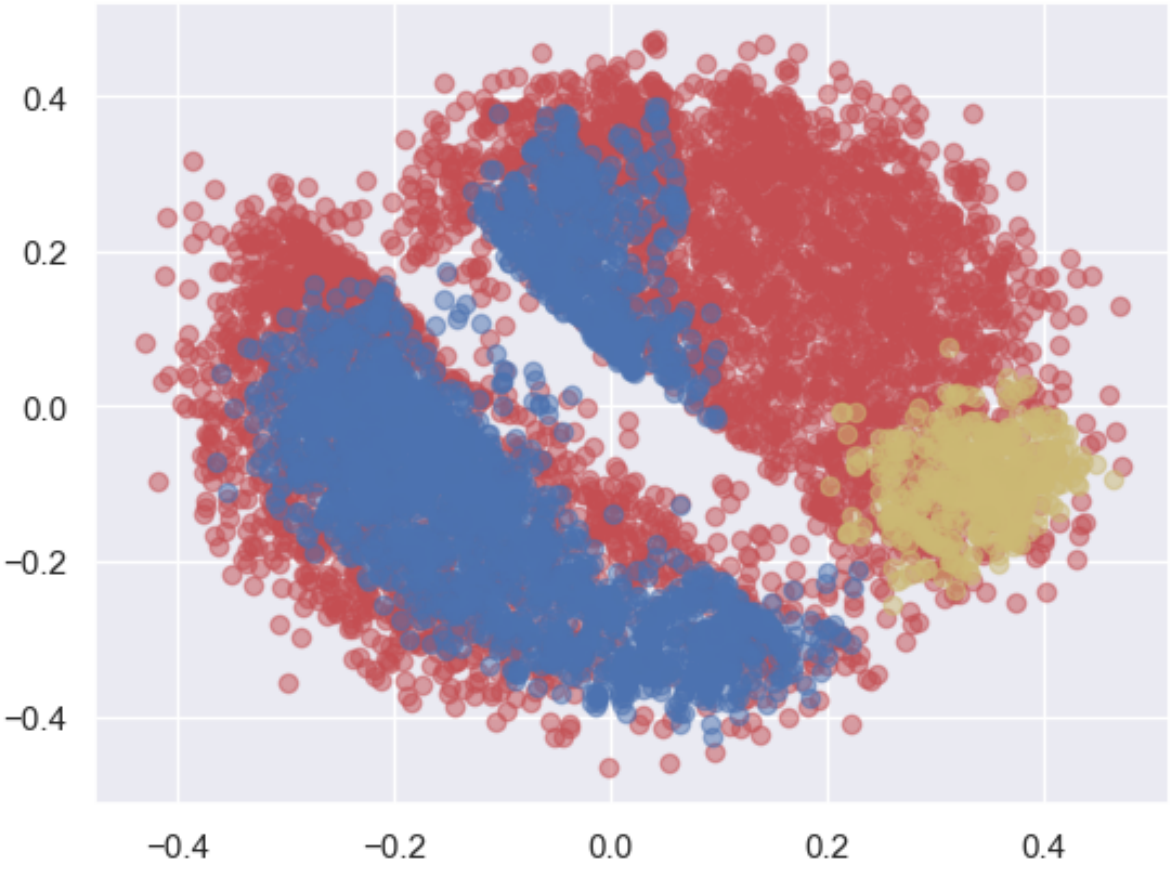

Figura 114: MDS pintat segons els clusters de DBSCAN

Els punts vermells són outliers i la resta de colors són clusters. Blau: c0 Groc: c1.

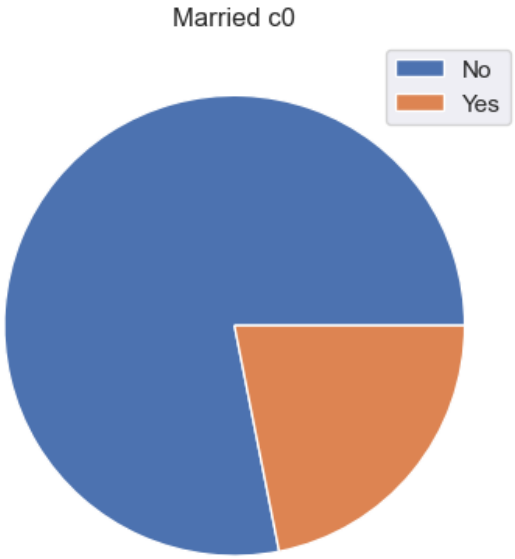

Figura 115: Pie plot Married c0

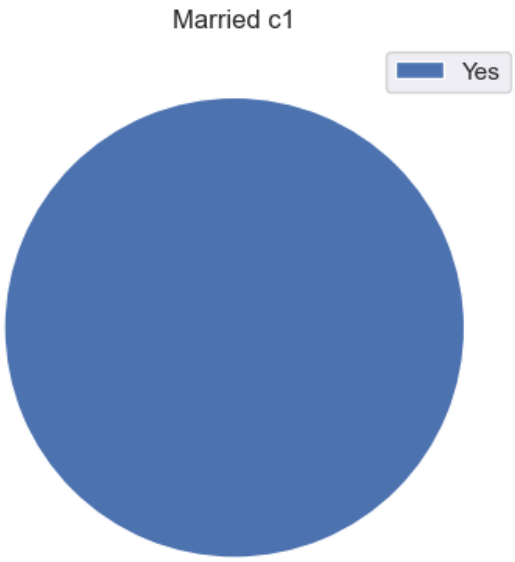

Figura 116: Pie plor Married c1

El primer cluster la majoria són solters mentre que en el segon estan tots casats.

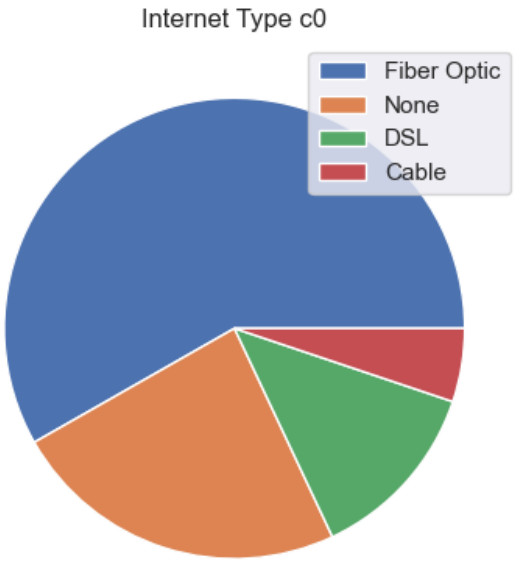

Figura 117: Pie plot Internet Type c0

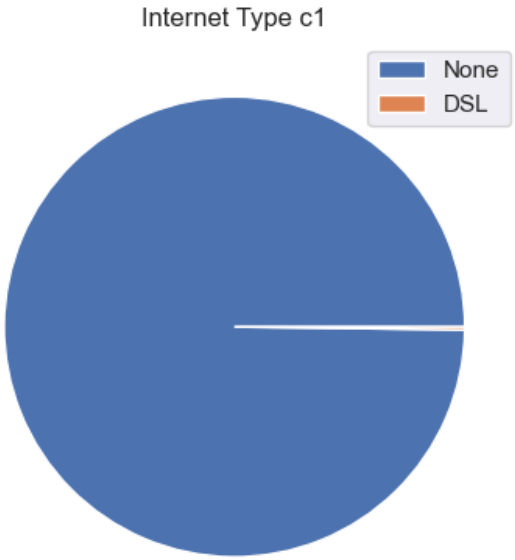

Figura 118: Pie plot Internet Type c1

En el primer cluster domina la fibra òptica mentre que en el 2n la majoria de clients no tenen ni internet contractat.

	Number of Dependents	Zip Code	Number of Referrals	Tenure in Months	Avg Monthly GB Download	Monthly Charge	Age_disc	Total Streaming	Premium Services	Refunds	Average Monthly Extra Data Charges	label
count	2933.000	2933.000	2933.000	2933.000	2933.000	2933.000	2933.000	2933.000	2933.000	2933.000	2933.000	2933.0
mean	0.048	0.576	0.208	0.263	0.194	0.469	0.473	0.319	0.299	0.028	0.008	0.0
std	0.215	0.297	0.406	0.285	0.198	0.298	0.275	0.399	0.238	0.164	0.043	0.0
min	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.0
25%	0.000	0.344	0.000	0.028	0.024	0.089	0.250	0.000	0.167	0.000	0.000	0.0
50%	0.000	0.574	0.000	0.155	0.165	0.559	0.500	0.000	0.333	0.000	0.000	0.0
75%	0.000	0.869	0.000	0.423	0.282	0.715	0.667	0.667	0.500	0.000	0.000	0.0
max	1.000	1.000	1.000	1.000	1.000	0.976	1.000	1.000	1.000	1.000	1.000	0.0

Figura 119: Describe c0

	Number of Dependents	Zip Code	Number of Referrals	Tenure in Months	Avg Monthly GB Download	Monthly Charge	Age_disc	Total Streaming	Premium Services	Refunds	Average Monthly Extra Data Charges	label
count	488.000	488.000	488.000	488.000	4.880e+02	488.000	488.000	488.0	488.000	488.000	488.0	488.0
mean	0.760	0.600	0.977	0.599	3.616e-04	0.047	0.385	0.0	0.055	0.037	0.0	1.0
std	0.427	0.303	0.149	0.324	7.988e-03	0.024	0.233	0.0	0.081	0.189	0.0	0.0
min	0.000	0.000	0.000	0.000	0.000e+00	0.016	0.000	0.0	0.000	0.000	0.0	1.0
25%	1.000	0.361	1.000	0.310	0.000e+00	0.029	0.167	0.0	0.000	0.000	0.0	1.0
50%	1.000	0.639	1.000	0.641	0.000e+00	0.034	0.417	0.0	0.000	0.000	0.0	1.0
75%	1.000	0.873	1.000	0.901	0.000e+00	0.074	0.583	0.0	0.167	0.000	0.0	1.0
max	1.000	1.000	1.000	1.000	1.765e-01	0.187	1.000	0.0	0.500	1.000	0.0	1.0

Figura 120: Describe c1

En el primer cluster (c0) els clients tenen més càrrecs mensuals, són més grans i tenen més serveis premium i de streaming contractats. En el 2n cluster cap client té serveis de streaming contractats, hi ha més proporció amb dependents en les seves cases, hi ha més proporció que han referit a amics o familiars, descarreguen més GB i porten més temps en la companyia.

9.2 Heriarchical

En aquest mètode hem de triar linkage per decidir com calcular la distància entre clusters i quina mètrica utilitzar per calcular distàncies.
En general ward és el linkage que millors resultats treu, per això decidim utilitzar-ho. Aquest es basa en minimitzar la variància. Pel jeràrquic també podríem utilitzar gower, però la implementació de sklearn no ens permet utilitzar ward a menys que utilitzem la distància euclídia, així que la utilitzem.
Fem el dendograma per decidir per on tallar:

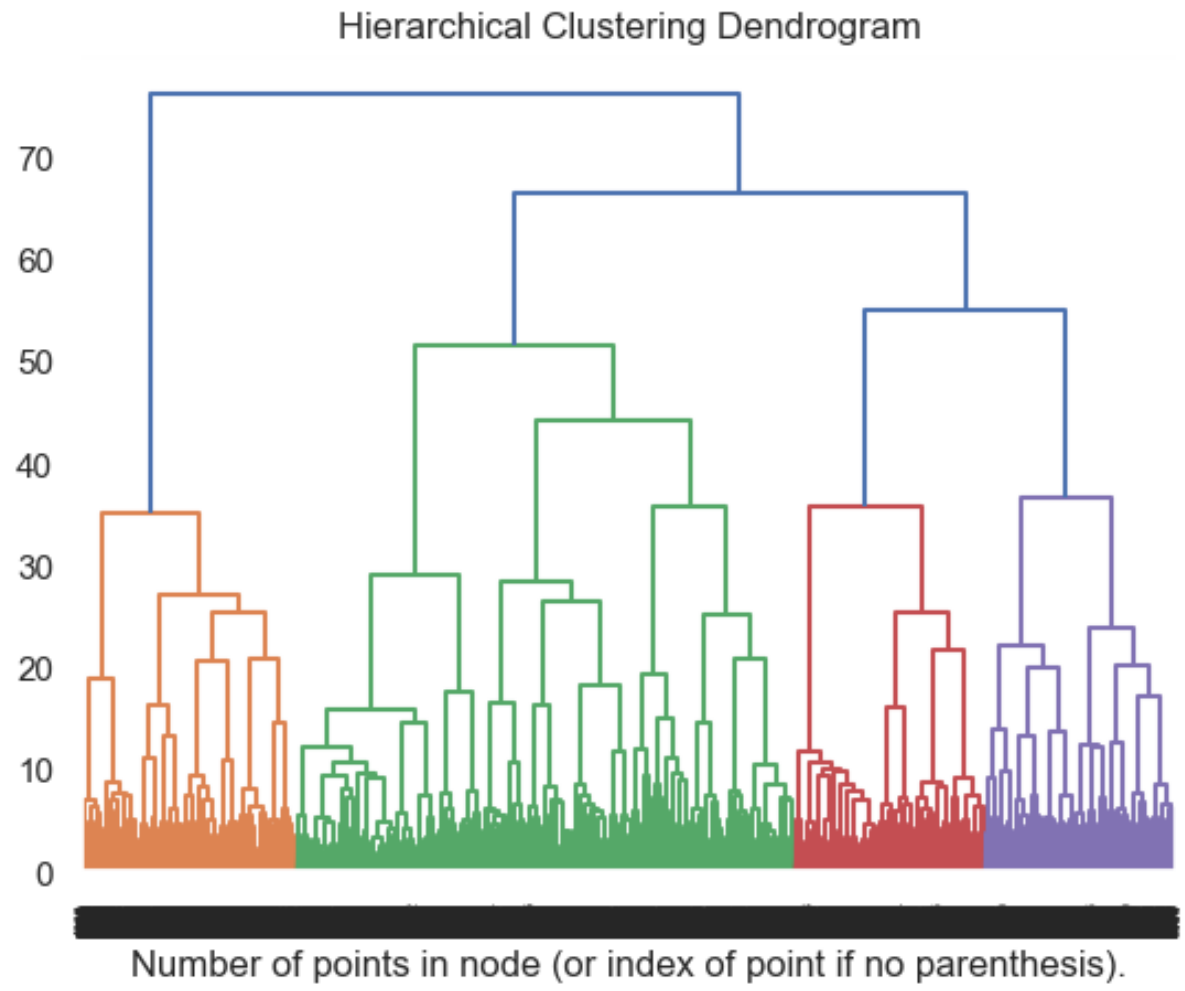

Figura 121: Dendograma

Els dos punts on es fa un split que estan més allunyats és entre 3 i 4 clusters, de forma que fem un tall en 3.

Interpretació resultats

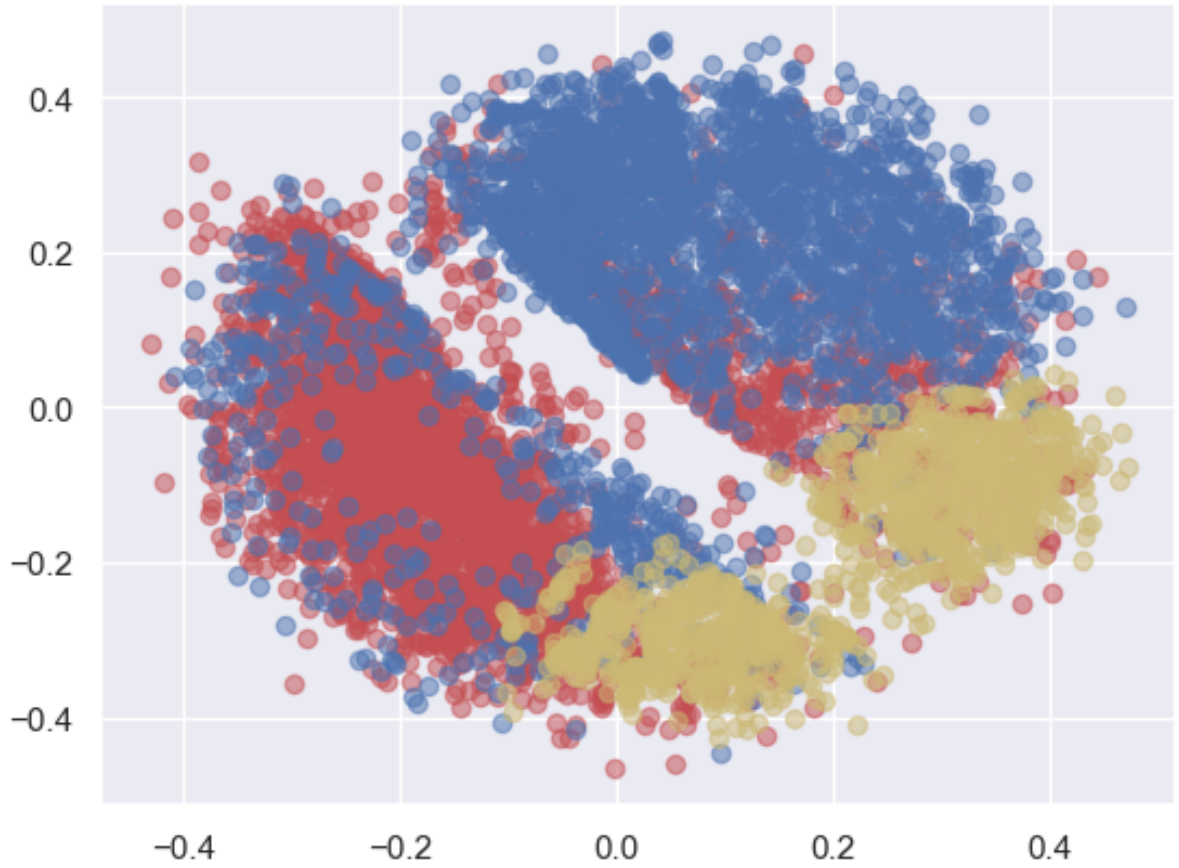

Figura 122: MDS pintat segons els colors de hierarchical clustering

Cada color és un cluster diferent: blau: c0, groc:c1, vermell: c2. En aquesta ocasió tampoc s'assemblen els clusters de MDS i els de clustering.
Respecte les variables categòriques, el cluster 0 té més representació de casats. Tots tenen internet i la majoria de clients paguen sense papers a través del banc. El cluster 1 cap client té internet, la majoria de clients reben els papers dels pagaments i paguen amb targeta.
El cluster 2 La majoria de clients està soltera. Paguen mes a mes i tots els que paguen per correu es troben en aquest cluster.

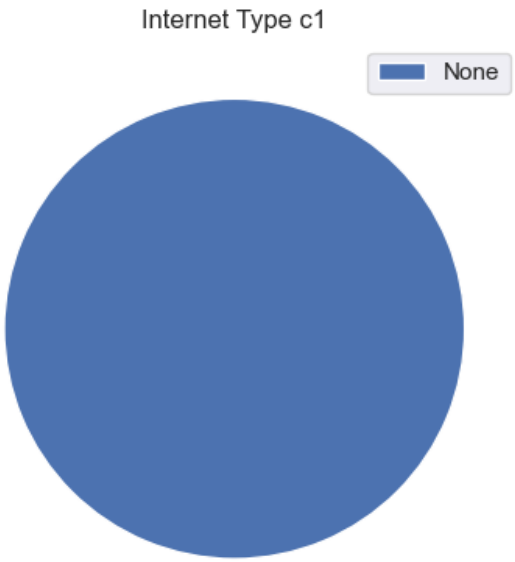

Figura 123: Pie plot Internet Type c1

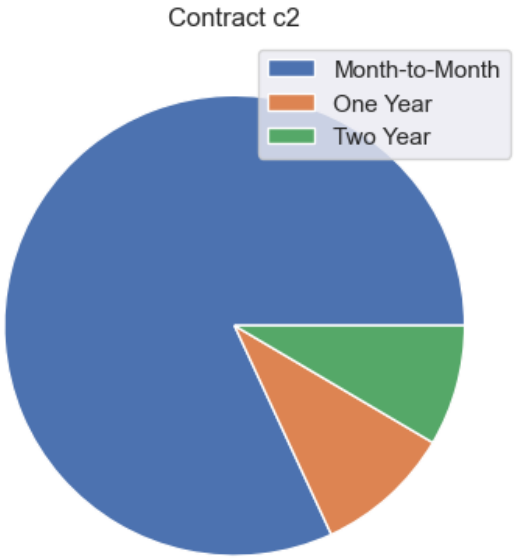

Figura 124: Pie plot Contract c2

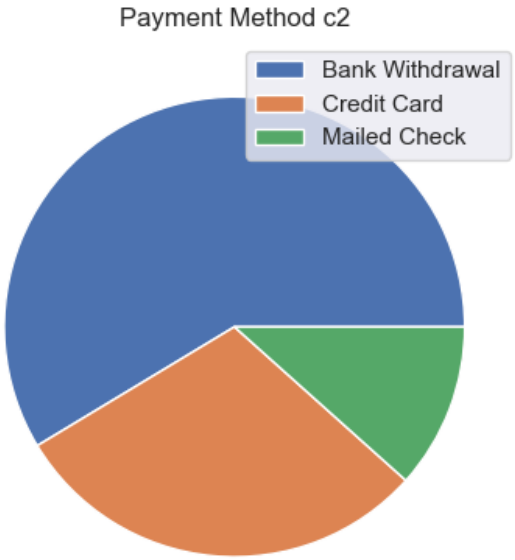

Figura 125: Pie plot Payment Method c2

	Number of Dependents	Zip Code	Number of Referrals	Tenure in Months	Avg Monthly GB Download	Monthly Charge	Age_disc	Total Streaming	Premium Services	Refunds	Average Monthly Extra Data Charges	label
count	2452.000	2452.000	2452.000	2452.000	2452.000	2452.000	2452.000	2452.000	2452.000	2452.000	2452.000	2452.0
mean	0.269	0.570	0.798	0.647	0.316	0.619	0.469	0.583	0.601	0.094	0.004	0.0
std	0.444	0.303	0.402	0.308	0.235	0.197	0.293	0.405	0.228	0.292	0.020	0.0
min	0.000	0.000	0.000	0.000	0.024	0.071	0.000	0.000	0.000	0.000	0.000	0.0
25%	0.000	0.344	1.000	0.394	0.153	0.472	0.250	0.333	0.500	0.000	0.000	0.0
50%	0.000	0.574	1.000	0.718	0.247	0.631	0.500	0.667	0.667	0.000	0.000	0.0
75%	1.000	0.869	1.000	0.930	0.353	0.779	0.667	1.000	0.833	0.000	0.000	0.0
max	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.625	0.0

Figura 126: Describe c0

	Number of Dependents	Zip Code	Number of Referrals	Tenure in Months	Avg Monthly GB Download	Monthly Charge	Age_disc	Total Streaming	Premium Services	Refunds	Average Monthly Extra Data Charges	label
count	1385.000	1385.000	1385.000	1385.000	1385.0	1385.000	1385.000	1385.0	1385.000	1385.000	1385.0	1385.0
mean	0.378	0.569	0.466	0.433	0.0	0.042	0.380	0.0	0.039	0.066	0.0	1.0
std	0.485	0.305	0.499	0.343	0.0	0.021	0.236	0.0	0.070	0.249	0.0	0.0
min	0.000	0.000	0.000	0.000	0.0	0.000	0.000	0.0	0.000	0.000	0.0	1.0
25%	0.000	0.328	0.000	0.113	0.0	0.028	0.167	0.0	0.000	0.000	0.0	1.0
50%	0.000	0.590	0.000	0.366	0.0	0.033	0.417	0.0	0.000	0.000	0.0	1.0
75%	1.000	0.869	1.000	0.746	0.0	0.041	0.583	0.0	0.000	0.000	0.0	1.0
max	1.000	1.000	1.000	1.000	0.0	0.098	1.000	0.0	0.167	1.000	0.0	1.0

Figura 127: Describe c1

	Number of Dependents	Zip Code	Number of Referrals	Tenure in Months	Avg Monthly GB Download	Monthly Charge	Age_disc	Total Streaming	Premium Services	Refunds	Average Monthly Extra Data Charges	label
count	3206.000	3206.000	3206.000	3206.000	3206.000	3206.000	3206.000	3206.000	3206.000	3206.000	3206.000	3206.0
mean	0.138	0.560	0.193	0.289	0.288	0.527	0.449	0.378	0.388	0.063	0.010	2.0
std	0.345	0.304	0.395	0.289	0.230	0.237	0.285	0.403	0.225	0.244	0.056	0.0
min	0.000	0.000	0.000	0.000	0.000	0.019	0.000	0.000	0.000	0.000	0.000	2.0
25%	0.000	0.328	0.000	0.042	0.129	0.330	0.250	0.000	0.167	0.000	0.000	2.0
50%	0.000	0.557	0.000	0.183	0.235	0.561	0.417	0.333	0.333	0.000	0.000	2.0
75%	0.000	0.869	0.000	0.479	0.341	0.712	0.667	0.667	0.500	0.000	0.000	2.0
max	1.000	1.000	1.000	1.000	1.000	0.984	1.000	1.000	1.000	1.000	1.000	2.0

Figura 128: Describe c2

Respecte les variables numèriques; el cluster 0 són clients més antics que tenen contractats més serveis de premium i streaming i per tant paguen més. També tenen més proporció de clients que han referit a algú. El cluster 1 són clients que no tenen internet contractat i que per tant no tenen serveis addicionals contractats apart d'alguns clients que tenen múltiples linies telefòniques. Al no tenir aquests serveis paguen molt menys que la resta. El cluster 3 està caracteritzat per clients que porten menys temps però si tenen algun servei extra contractat. Com porten menys temps la proporció de clients que han referit a algú és menor.