

Predicción de fuga de clientes - TELCO

Ricardo Bautista Huerta

Problemática de negocio

El presente caso es una empresa de telecomunicaciones que ofrece servicios, incluyendo telefonía fija y móvil e internet.

Una alta cantidad de clientes están cancelando sus servicios y contratando servicios de la competencia por varias razones como: insatisfacción por el servicio, costos, mejor servicio de la competencia, actitud del soporte, etc.

Entonces, por lo que se requiere:

- Aumentar la retención de clientes
- Reducir la tasa de fuga de clientes
- Mejorar la satisfacción de los clientes



Problemática de negocio

- La fuga de clientes es costosa:

- Pérdidas de ingreso y costos de adquisición

"[...] Atraer nuevos clientes resulta unas 6 o 7 veces más costoso"

- Genera una mala imagen de la empresa

- Mala percepción y reputación

- Saber qué se está haciendo mal y en qué se puede mejorar



Base de datos

Perteneciente a IBM como parte de bases de ejemplo de IBM Cognos Analytics

11.1.3: <https://www.kaggle.com/datasets/ylchang/telco-customer-churn-1113/data>

Conjunto de diferentes tablas con información de 7043 clientes de una empresa de telecomunicaciones que da servicios de teléfono e Internet a domicilio en California durante el Tercer Trimestre (Q3)

Compuesto por 5 tablas:

1 | Demografía

Código del Cliente, Género, Edad, Casado, Dependientes

2 | Ubicación

Código del Cliente, Estado, Ciudad, Código Zip, Latitud, Longitud

3 | Población

Código Zip, Población

4 | Servicios

Código del Cliente, Recomendó a un amigo, Número de referencias, Permanencia en meses, Tipo de Oferta, Múltiples líneas, Servicio de internet, Seguridad online, Backup Online, Cargos totales, Reembolsos totales, etc.

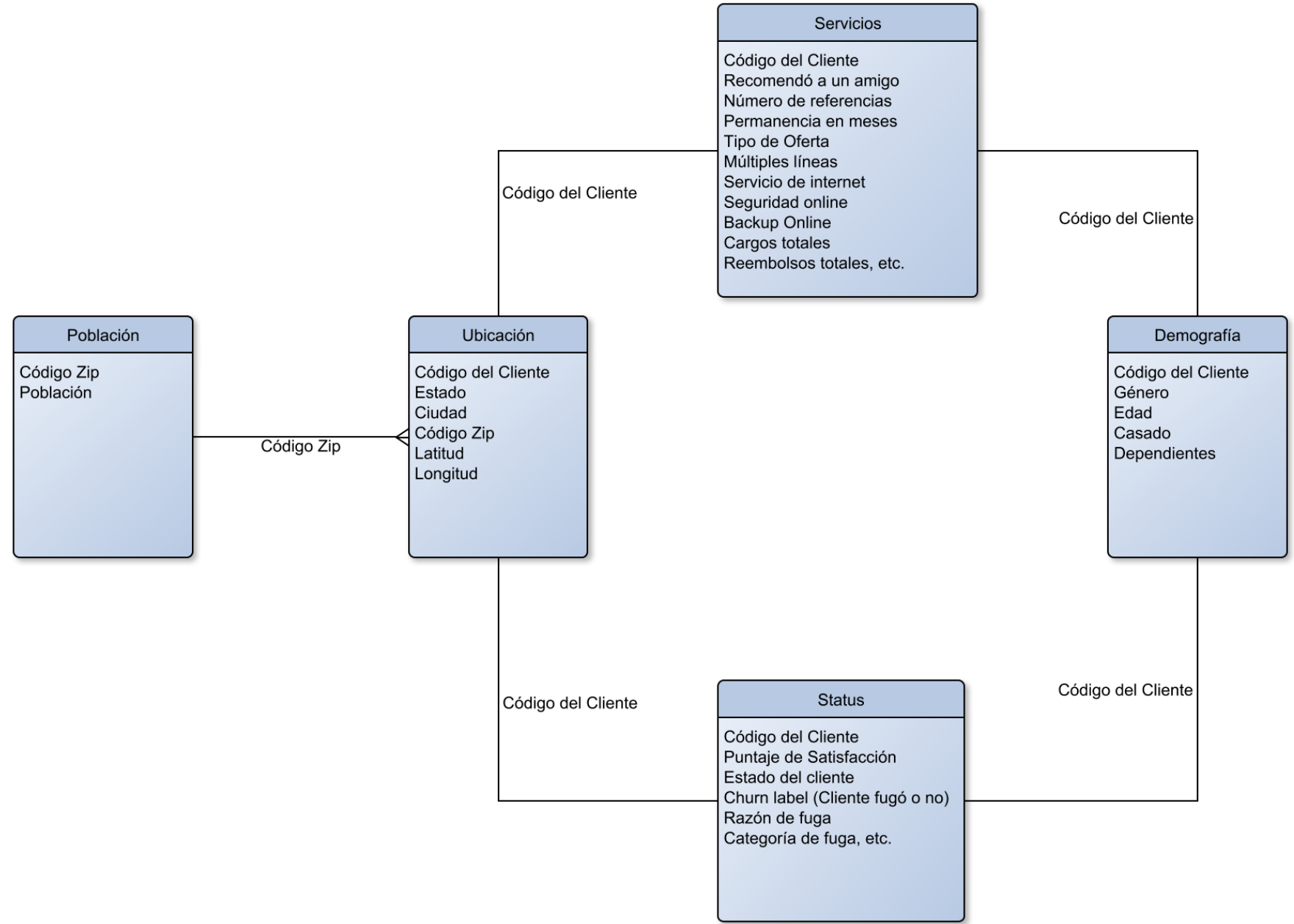
5 | Status

Código del Cliente, Puntaje de Satisfacción, Estado del cliente, Churn label (Cliente fugó o no), Razón de fuga, Categoría de fuga, etc.

Base de datos

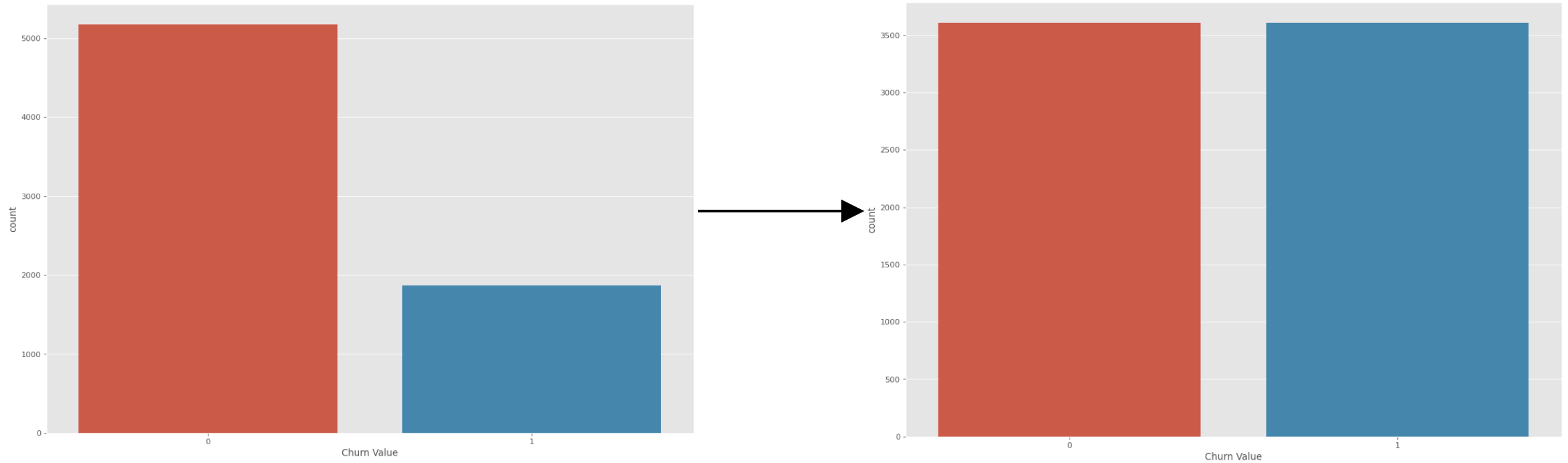
Se hizo merge a las bases para resultar en una tabla final.

Cada merge se hizo con base en las variables identificadoras.



Modelado

Primero, se hizo una combinación de under over sampling utilizando SMOTETomek de imbalanced-learn debido al fuerte desbalance de la base de datos



Modelado

Se probaron 9 modelos base usando cross validation con 10 k-folds para determinar cuál se comporta mejor utilizando como métrica el f1 score:

KNN

Decision Tree

Bagging Trees

Logistic Regression (L1) y (L2)

Random Forest

AdaBoost Classifier

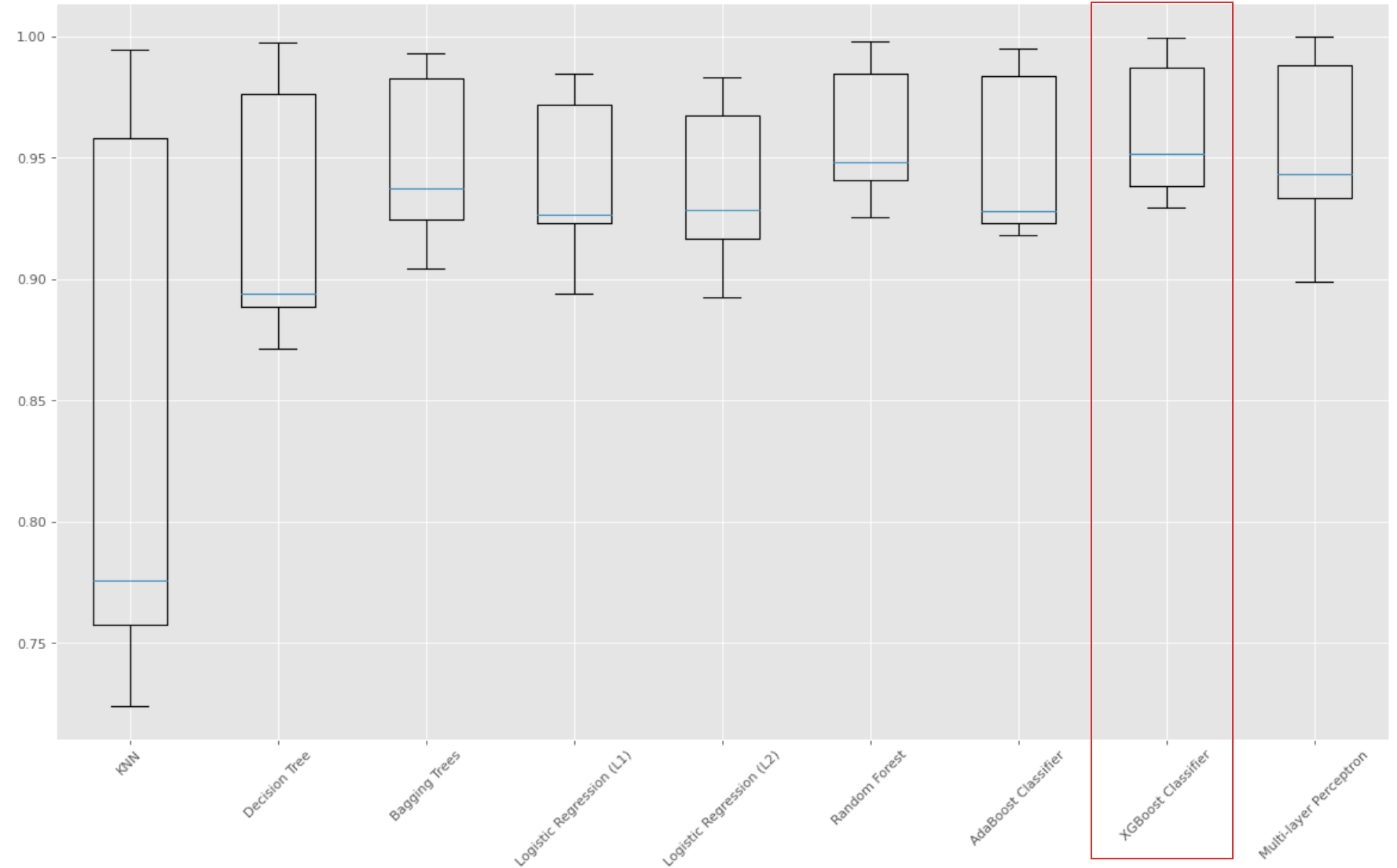
XGBoost Classifier

Multi-layer Perceptron

```
1 num_folds = 10
2
3
4 models = []
5 models.append(("KNN", KNeighborsClassifier()))
6 models.append(("Decision Tree", DecisionTreeClassifier(max_depth=5)))
7 models.append(("Bagging Trees", BaggingClassifier()))
8 models.append(("Logistic Regression (L1)", LogisticRegression(penalty='l1', solver='liblinear')))
9 models.append(("Logistic Regression (L2)", LogisticRegression(penalty='l2')))
10 models.append(("Random Forest", RandomForestClassifier(n_estimators=200, criterion='entropy')))
11 models.append(("AdaBoost Classifier", AdaBoostClassifier(n_estimators=200)))
12 models.append(("XGBoost Classifier", xgb.XGBClassifier(n_estimators=200, objective='binary:logistic')))
13 models.append(("Multi-layer Perceptron", MLPClassifier(solver='lbfgs')))
```

Modelado

Comparación de algoritmos - f1

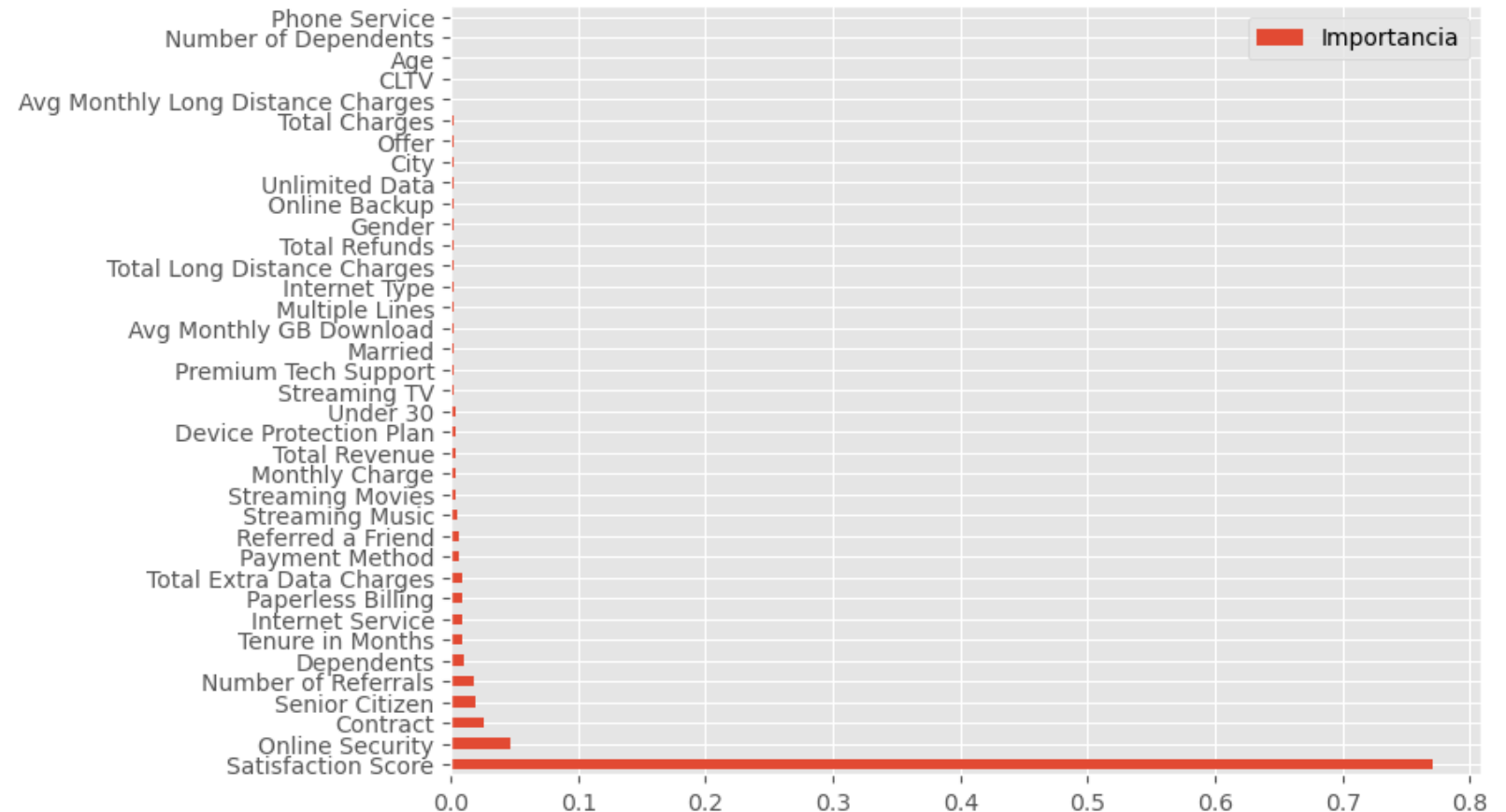


KNN: 0.838 (0.108)
Decision Tree: 0.922 (0.050)
Bagging Trees: 0.948 (0.032)
Logistic Regression (L1): 0.939 (0.031)
Logistic Regression (L2): 0.937 (0.031)
Random Forest: 0.958 (0.026)
AdaBoost Classifier: 0.948 (0.032)
XGBoost Classifier: 0.960 (0.027)
Multi-layer Perceptron: 0.951 (0.036)

Modelado

Se seleccionaron las 17 variables más importantes dadas por el **XGBClassifier**:

Satisfaction Score
Online Security
Contract
Senior Citizen
Number of Referrals
Dependents
Tenure in Months
Internet Service
Paperless Billing
Total Extra Data Charges
Payment Method
Referred a Friend
Streaming Music
Streaming Movies,
Monthly Charge
Total Revenue
Device Protection Plan



Modelado

Se Tuneó el XGboost con HYPEROPT usando como métrica el 'AUC' para obtener los mejores hiperparámetros

```
1 space={'max_depth': hp.guniform("max_depth", 3, 18, 1),
2       'gamma': hp.uniform('gamma', 1,9),
3       'reg_alpha' : hp.guniform('reg_alpha', 40,180,1),
4       'reg_lambda' : hp.uniform('reg_lambda', 0,1),
5       'colsample_bytree' : hp.uniform('colsample_bytree', 0.5,1),
6       'min_child_weight': hp.guniform('min_child_weight', 0, 10, 1),
7       'n_estimators': 180,
8       'seed': 0,
9       'random_state':42
10      }
```

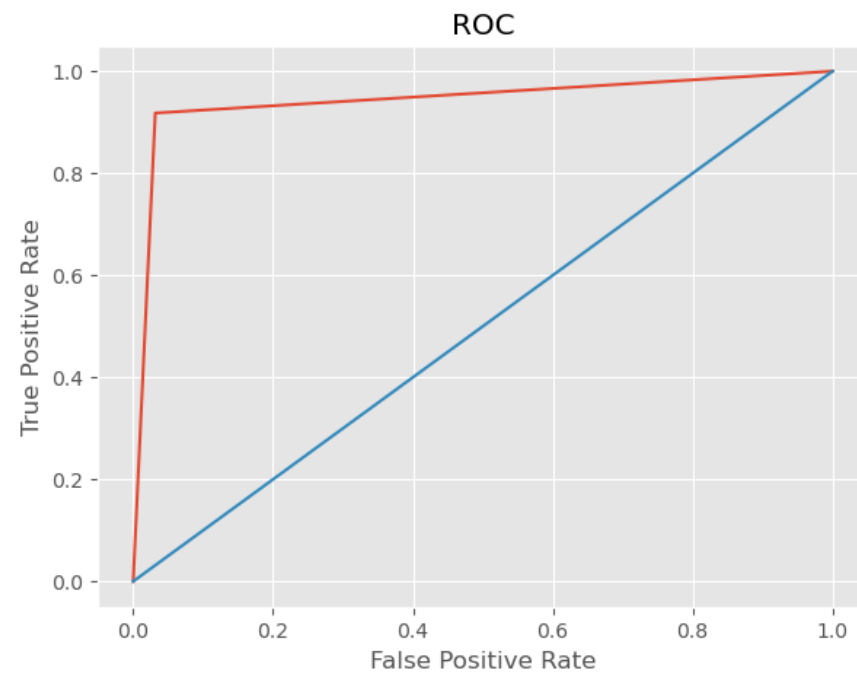
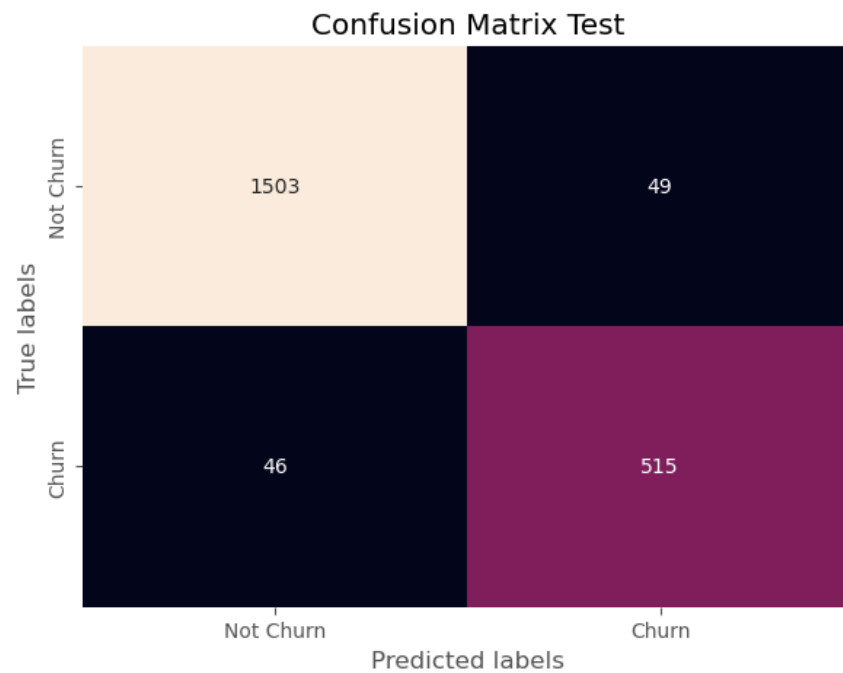
Mejores hyperparametros :

```
{'colsample_bytree': 0.8224506395273834, 'gamma': 7.523889936788071, 'max_depth': 4.0, 'min_child_weight': 10.0, 'reg_alpha': 41.0, 'reg_lambda': 0.6396149280329773}
```

Métricas

Se obtuvo:
accuracy de testing de 0.96

AUC ROC de 0.943



	precision	recall	f1-score	support
0	0.97	0.97	0.97	1552
1	0.91	0.92	0.92	561
accuracy			0.96	2113
macro avg	0.94	0.94	0.94	2113
weighted avg	0.96	0.96	0.96	2113

Deployment

Se utilizó Streamlit para desarrollar la aplicación que utilizará el modelo donde la predicción podrá ser online o subiendo un archivo csv



```
import pickle
import streamlit as st
import pandas as pd
from sklearn.preprocessing import StandardScaler
from PIL import Image

model_file = 'mejor_modelo.pkl'

model = pickle.load(open(model_file, 'rb'))
X_scaler_train = pd.read_csv('X_scaler_parameters.csv')

def main():

    # image = Image.open('images/icone.png')
    image2 = Image.open('images/image.png')
    # st.image(image, use_column_width=False)
    add_selectbox = st.sidebar.selectbox(
        "How would you like to predict?",
        ("Online", "Batch"))
    st.sidebar.info('This app is created to predict Customer Churn')
    st.sidebar.image(image2)
    st.title("Customer Churn Prediction - TELCO")
    if add_selectbox == 'Online':
        Satisfaction_Score = st.number_input(' Satisfaction Score : ', min_value=1, max_value=5, value=3)
        Online_Security = st.selectbox(' Online Security: ', ['Yes', 'No'])
        Contract = st.selectbox(' Contrato : ', ['Month-to-Month', 'One Year', 'Two Year'])
        Senior_Citizen = st.selectbox(' Senior Citizen : ', ['Yes', 'No'])
        Number_of_Referrals = st.number_input(' Number of Referrals: ', min_value=0, max_value=11, value=0)
        Dependents = st.selectbox(' Dependents: ', ['Yes', 'No'])
        Tenure_in_Months = st.number_input(' Tenure in Months : ', min_value=1, max_value=100, value=2)
        Internet_Service = st.selectbox(' Customer has Internet Service:', ['Yes', 'No'])
        Paperless_Billing = st.selectbox(' Customer has a Paperless Billing:', ['Yes', 'No'])
        Total_Extra_Data_Charges = st.number_input(' Total Extra Data Charges :', min_value=1, max_value=100, value=1)
        paymentmethod = st.selectbox('Payment Method:', ['Bank Withdrawal', 'Credit Card', 'Mailed Check'])
        Referred_a_Friend = st.selectbox(' Referred a Friend : ', ['Yes', 'No'])
        Streaming_Music = st.selectbox(' Streaming Music : ', ['Yes', 'No'])
        Streaming_Movies = st.selectbox(' Streaming Movies : ', ['Yes', 'No'])
        Monthly_Charge = st.slider(' Monthly Charge : ', 0, 120, 25)
        Total_Revenue = st.slider(' Total Revenue : ', 0, 20000, 1000)
        Device_Protection_Plan = st.selectbox(' Device Protection Plan : ', ['Yes', 'No'])
        output = ""
        output_prob = ""
```


Deployment

×

How would you like to predict?

Online

This app is created to predict Customer Churn



Customer Churn Prediction - TELCO

Satisfaction Score :

2

-

+

Online Security:

Yes

▼

Contrato :

Month-to-Month

▼

Senior Citizen :

Yes

▼

Number of Referrals:

0

-

+

Dependents:

Yes

▼

Tenure in Months :

2

-

+

Customer has Internet Service:

Yes

▼

Monthly Charge :

0

74

120

Total Revenue :

0

7395

20000

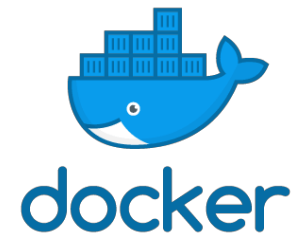
Device Protection Plan :

Yes

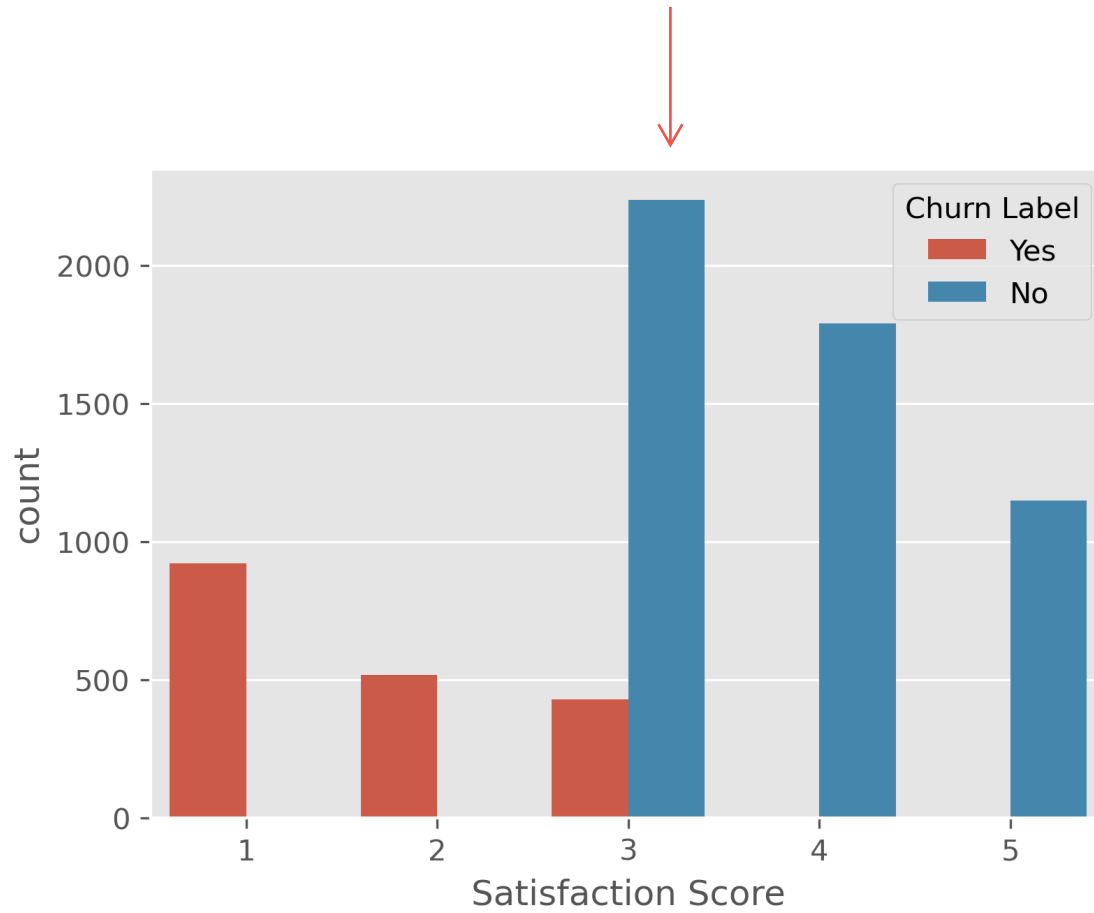
▼

Predict

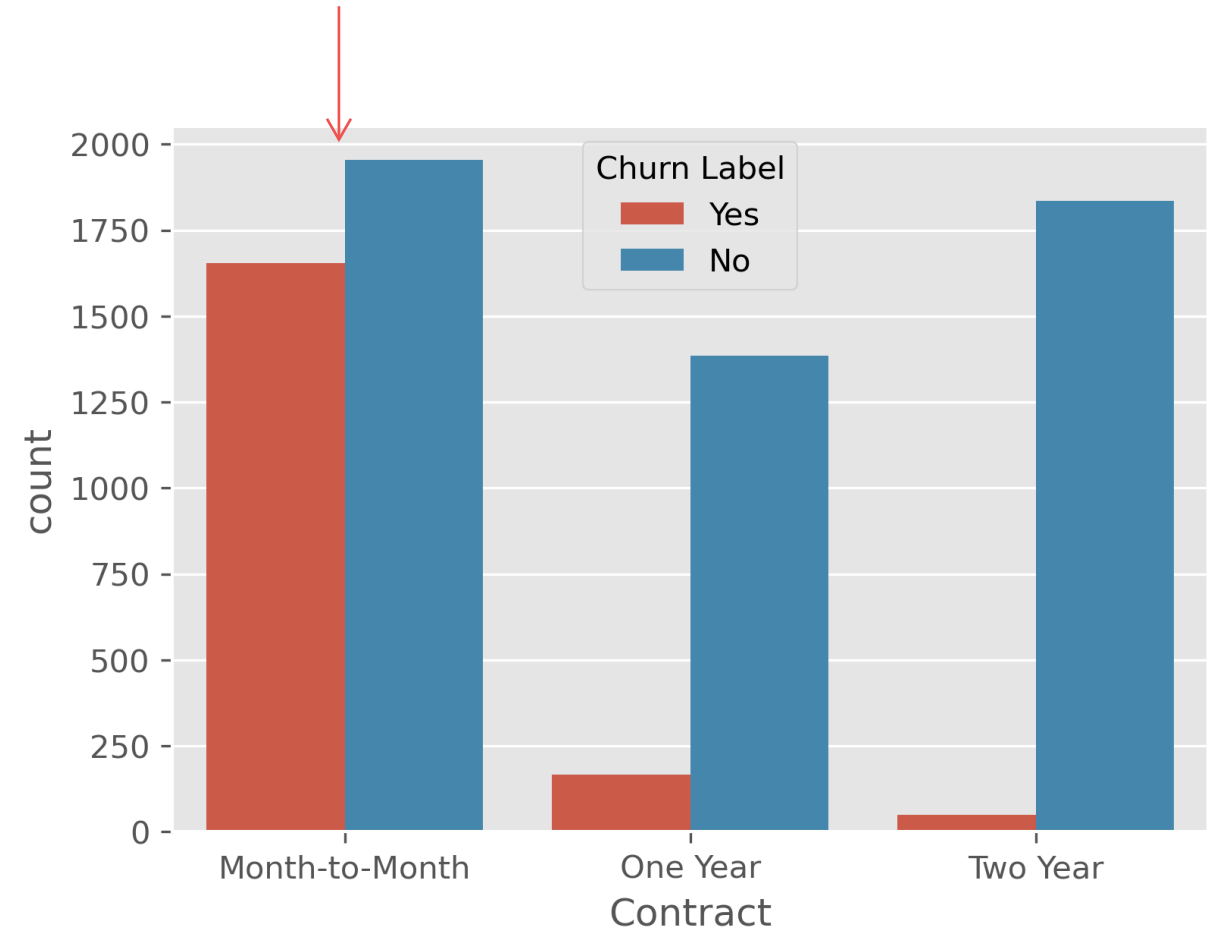
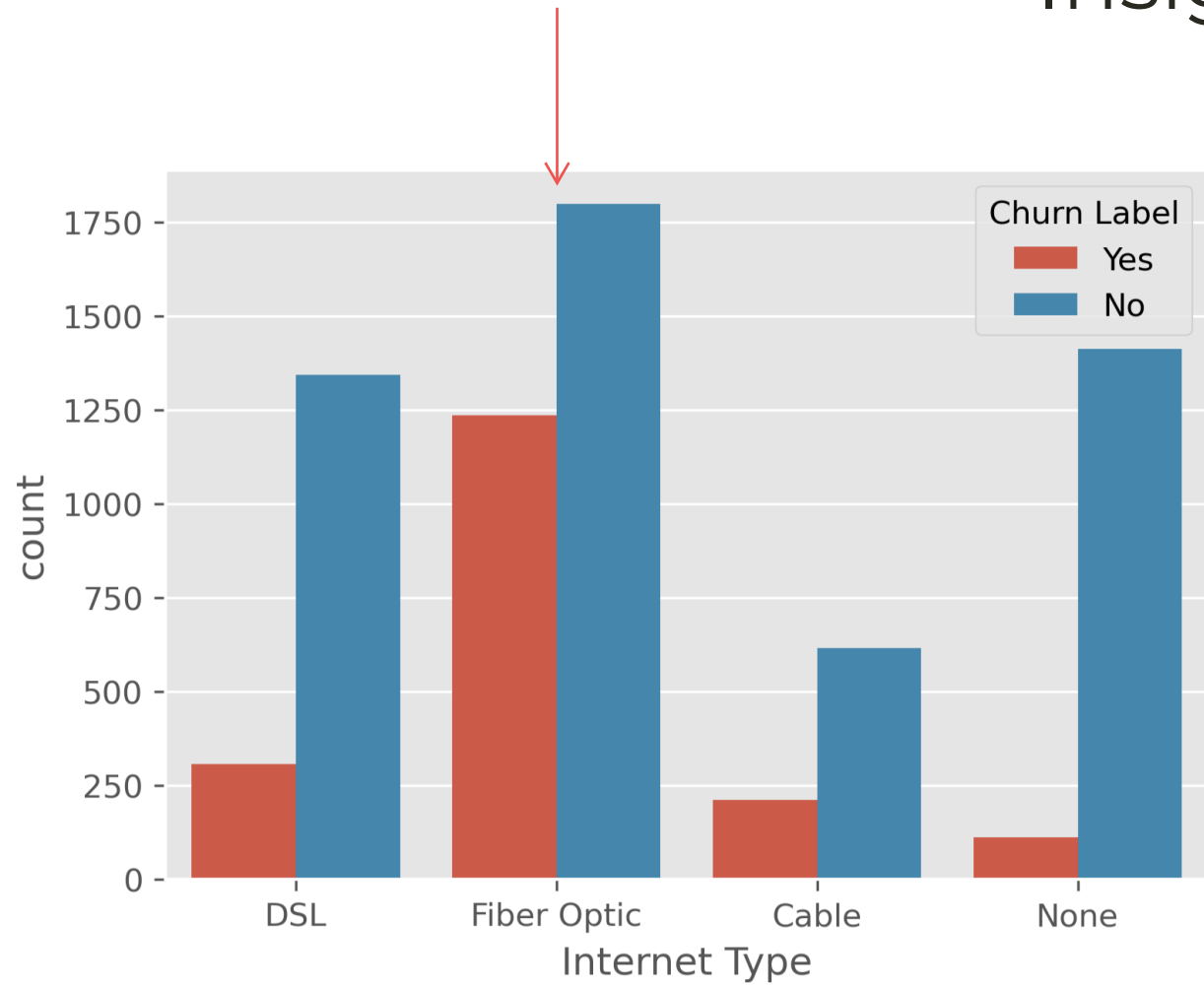
Churn: True, Risk Score: 0.979592502117157



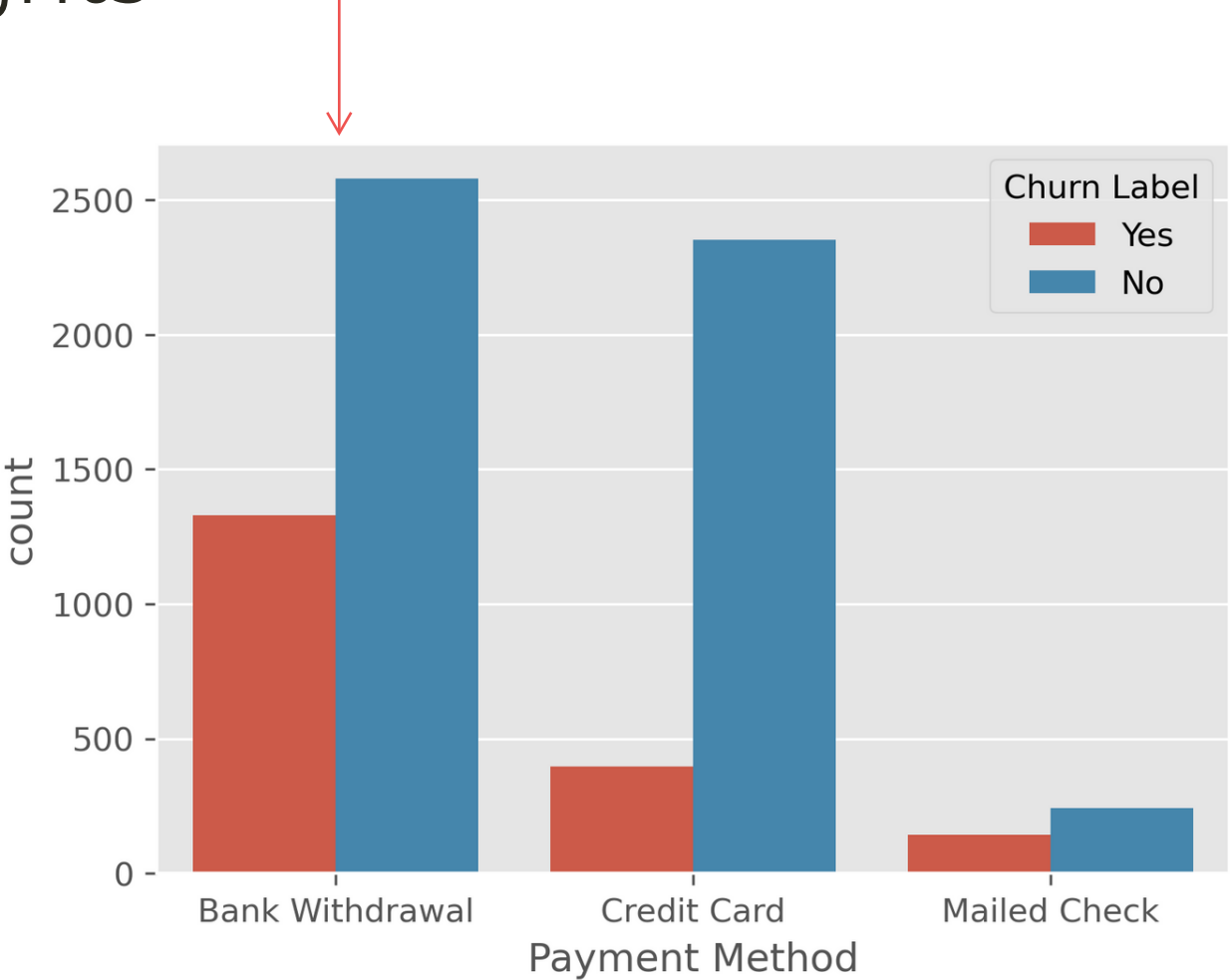
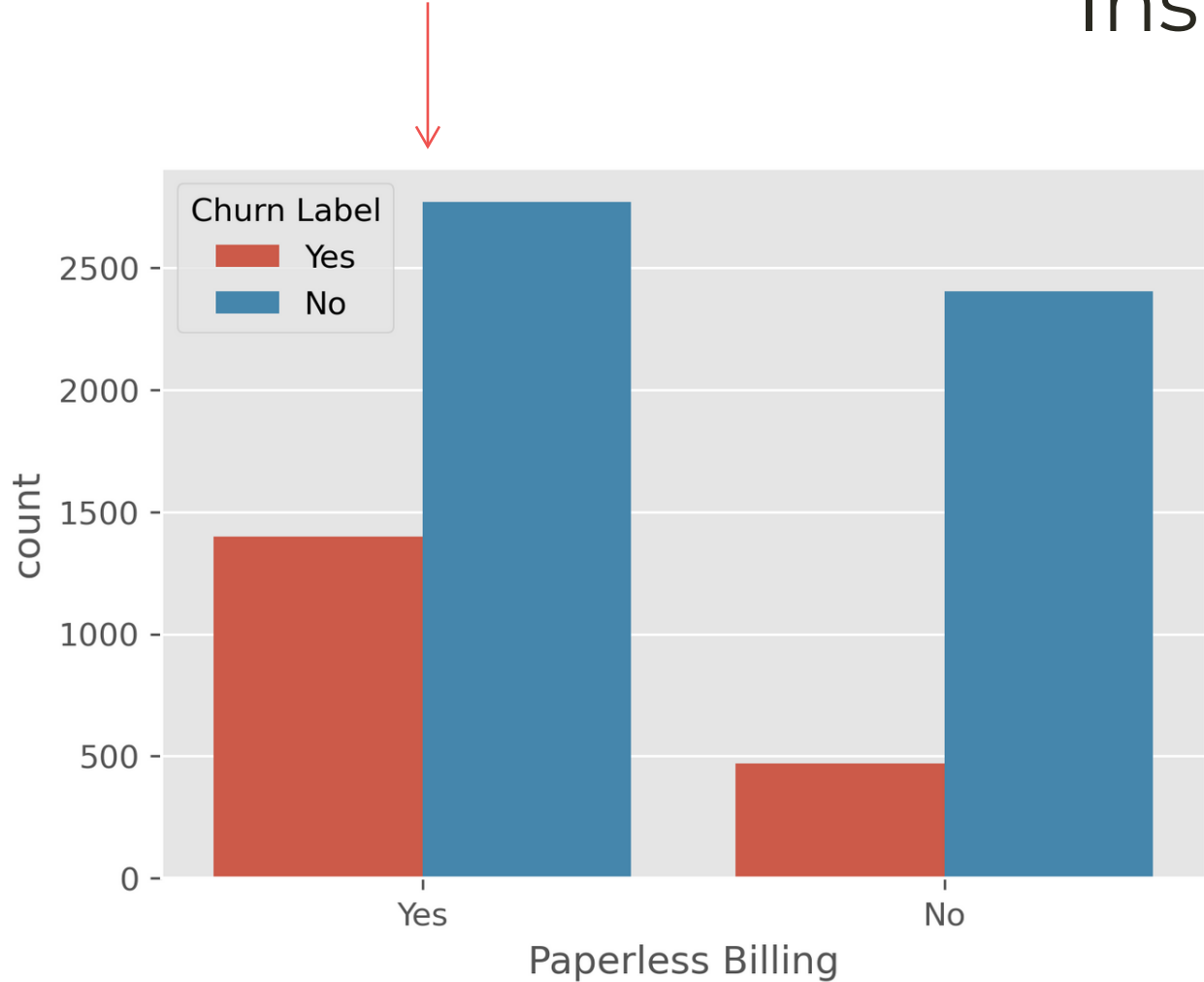
Insights



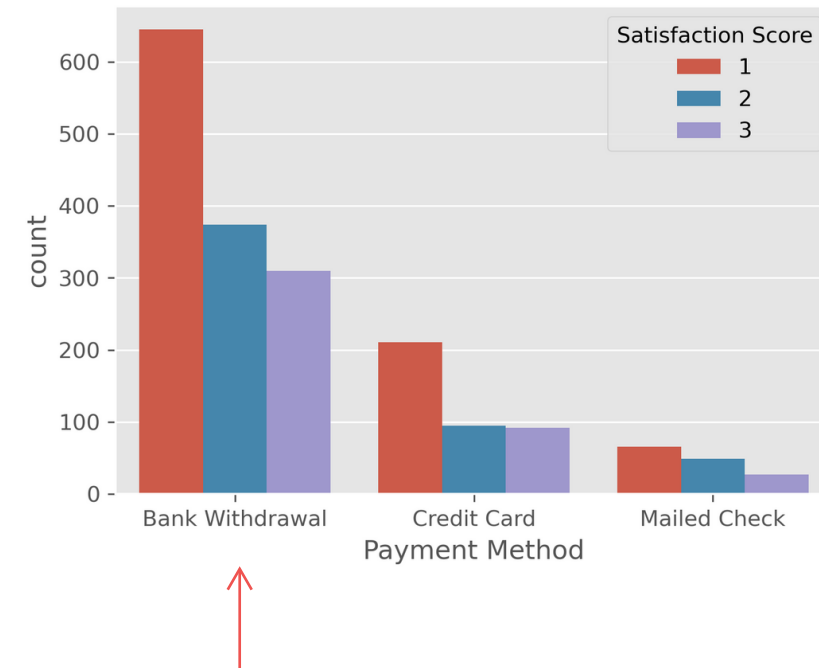
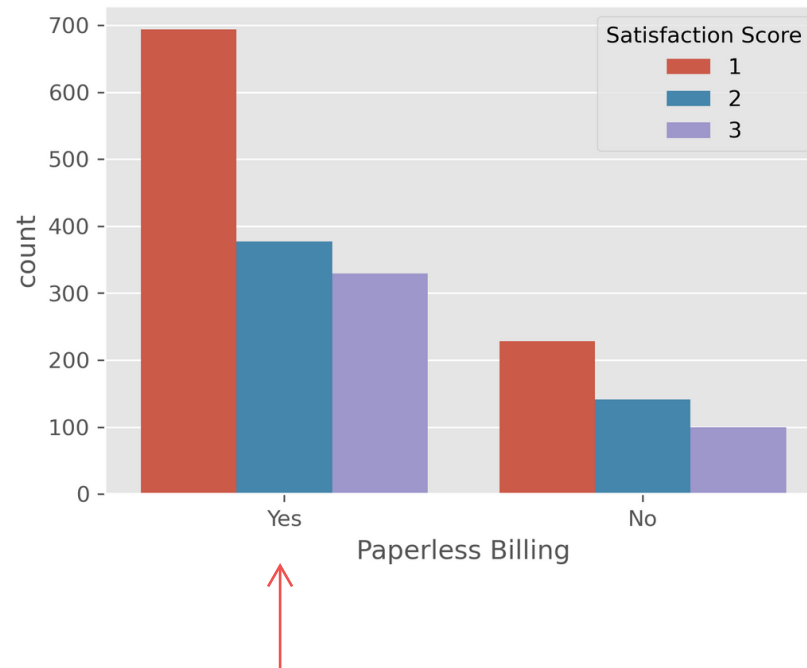
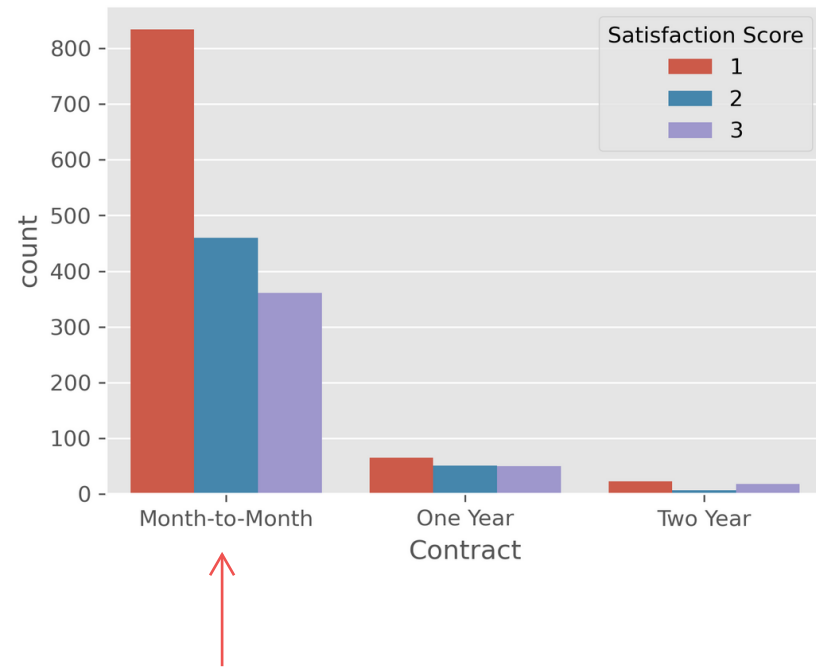
Insights



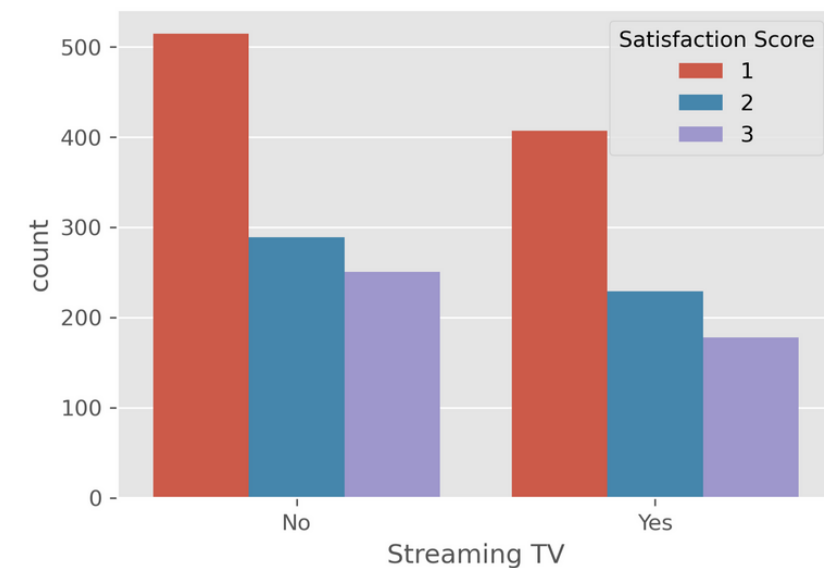
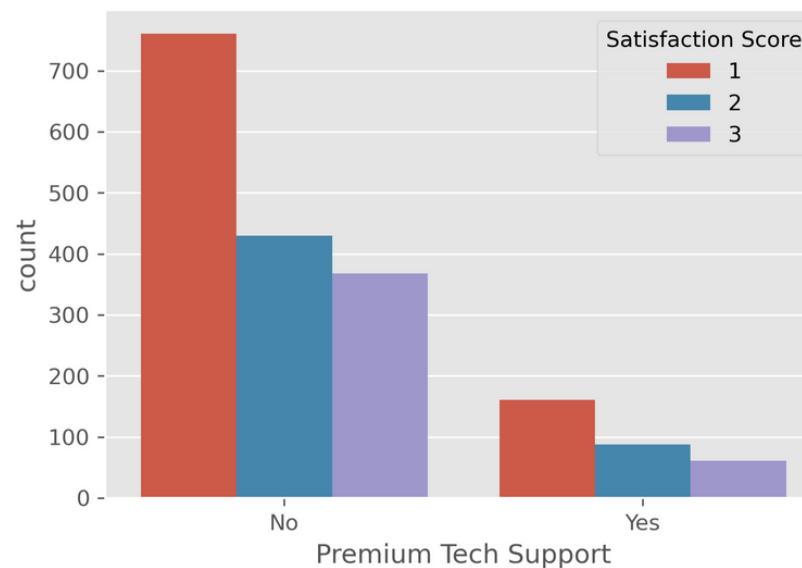
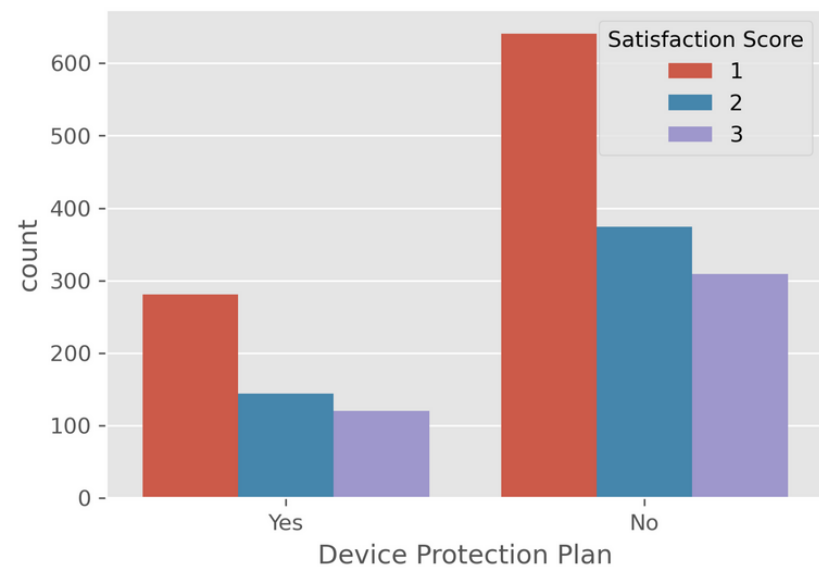
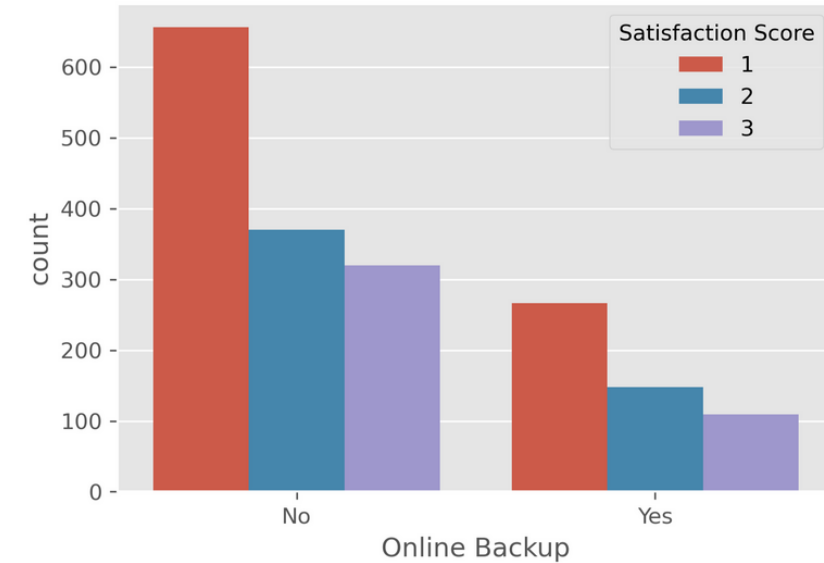
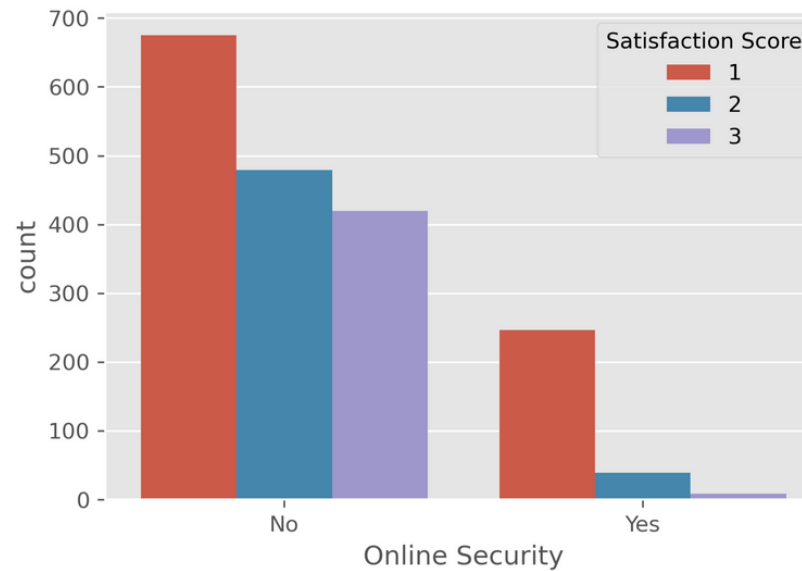
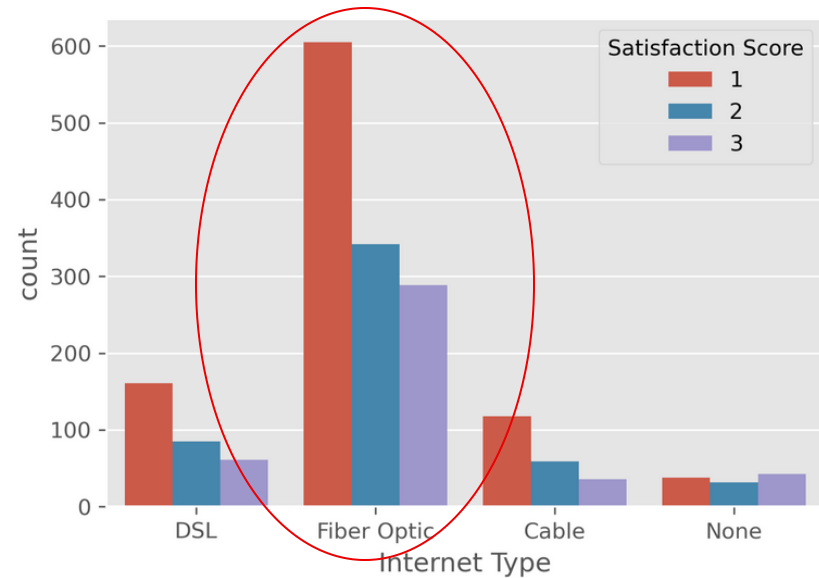
Insights



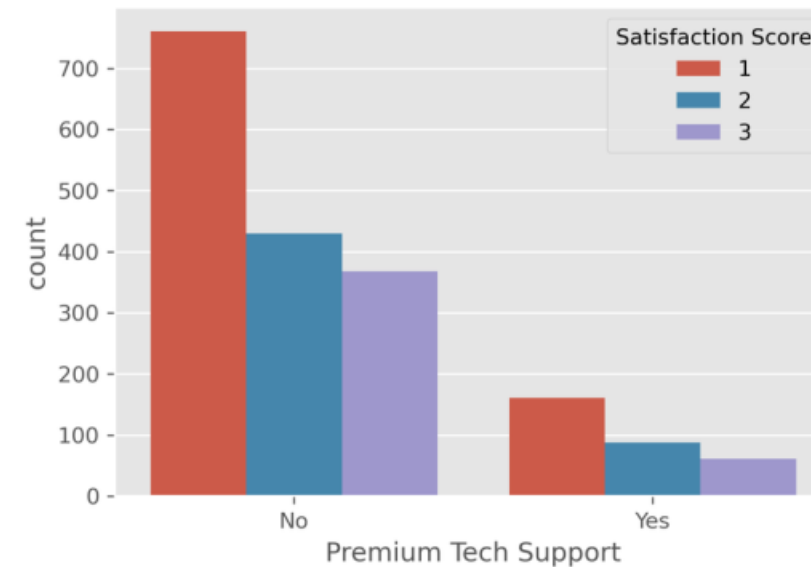
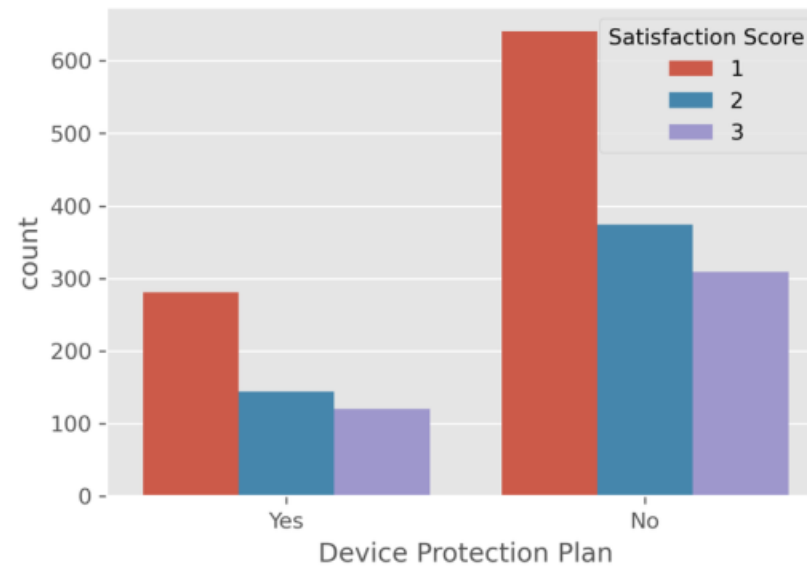
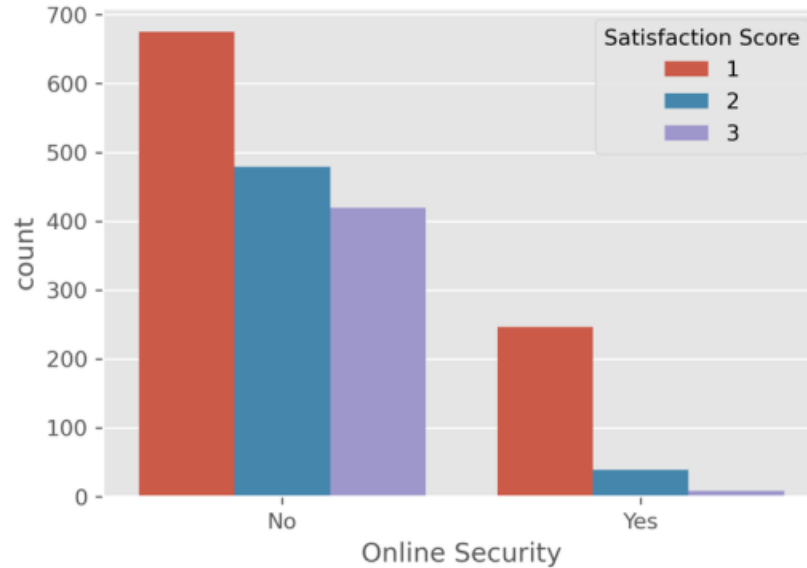
Insights - Solo Clientes que Abandonaron



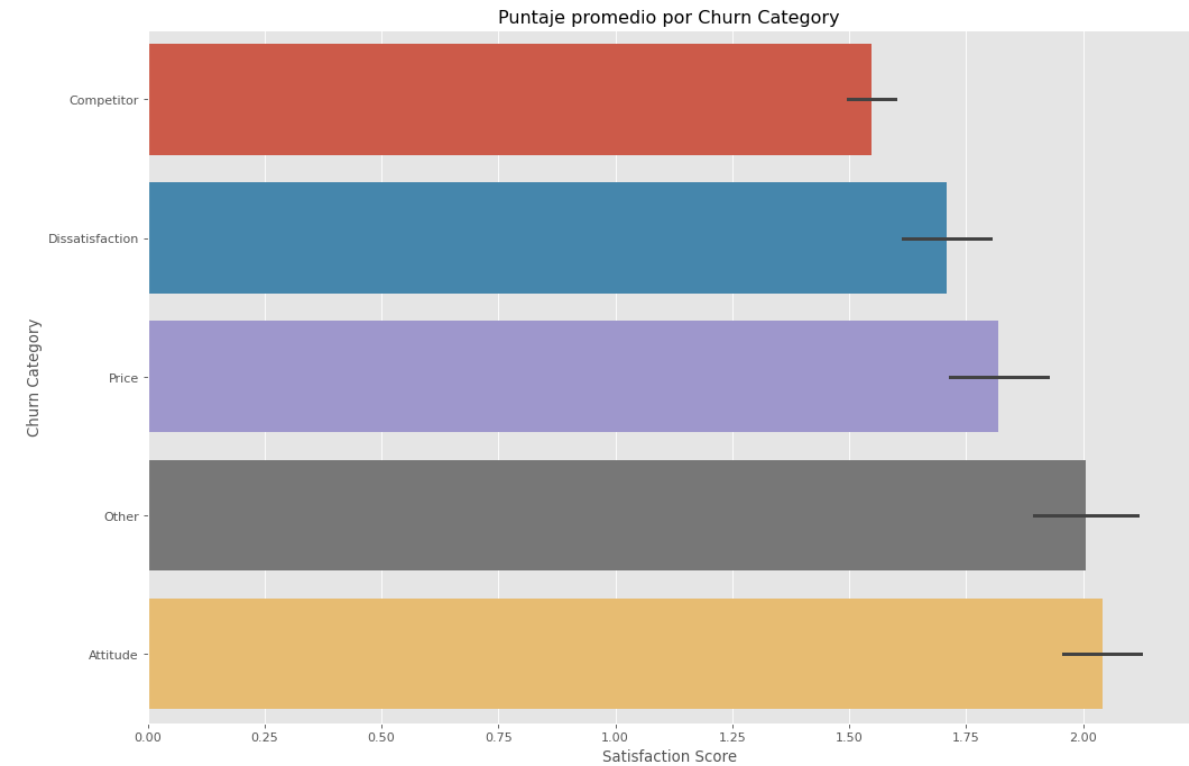
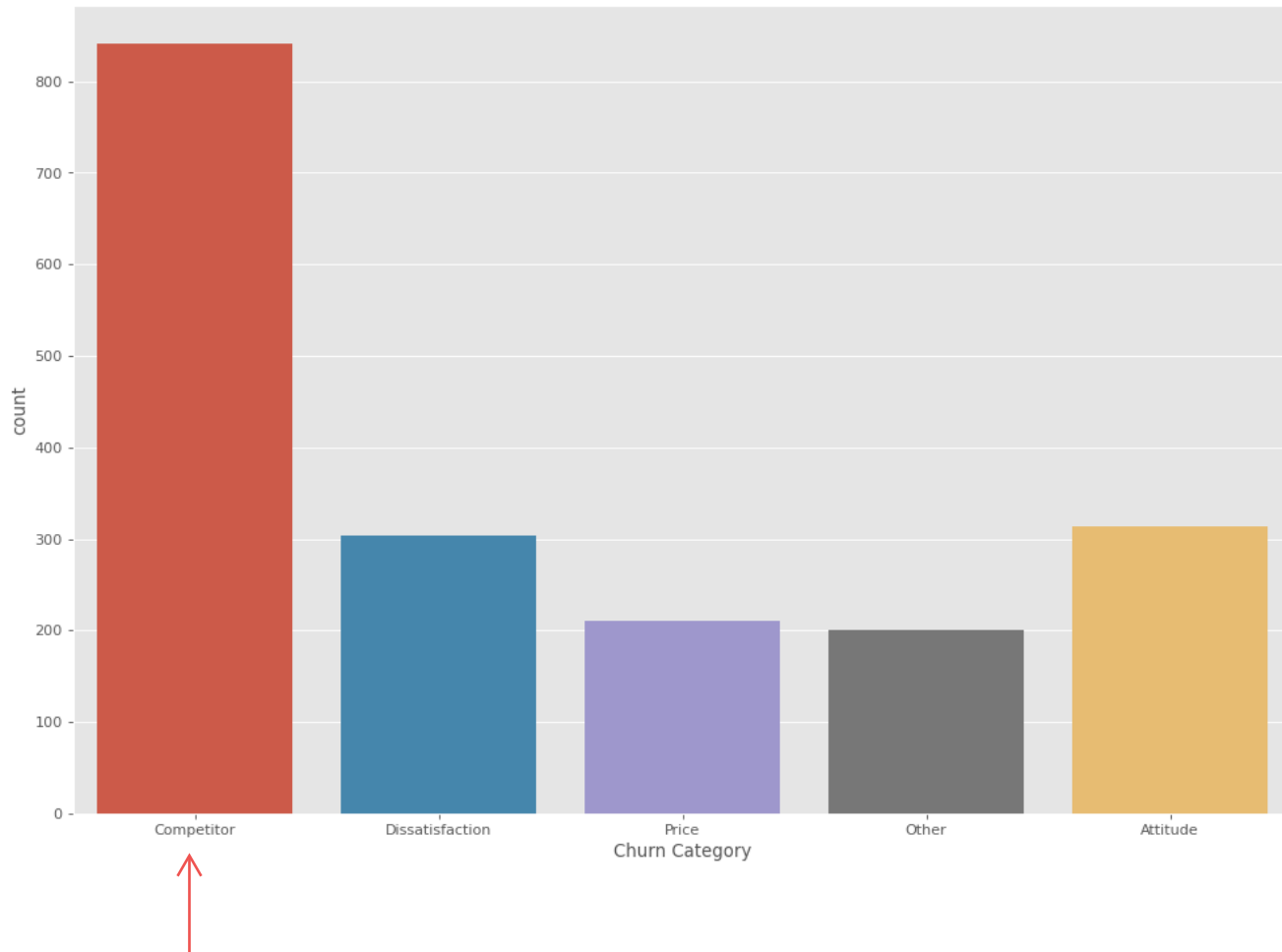
Insights - Solo Clientes que Abandonaron



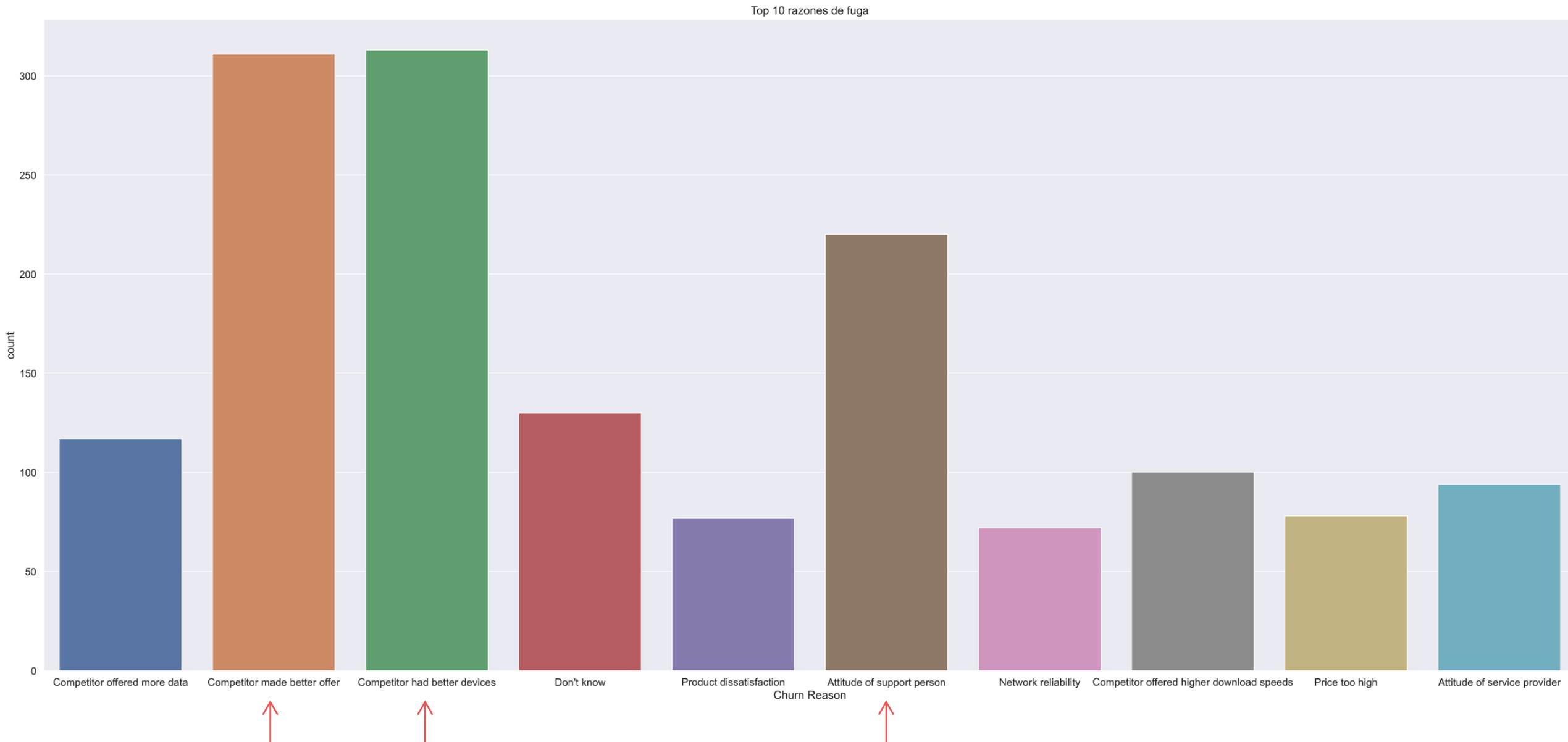
Insights - Solo Clientes que Abandonaron



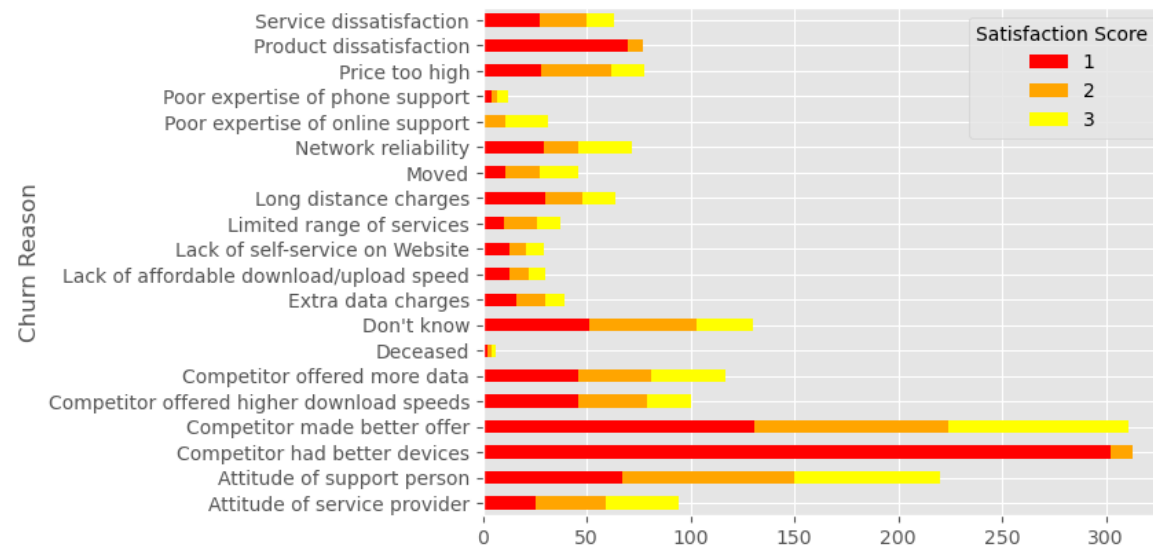
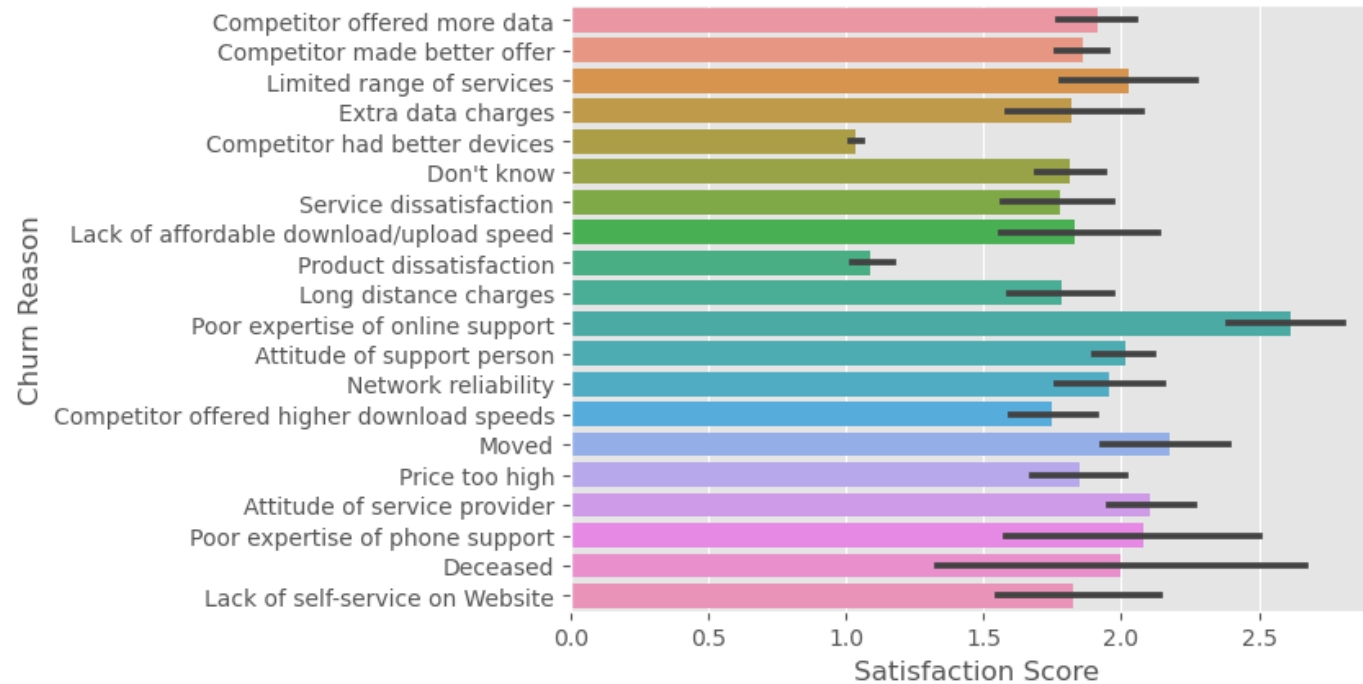
Insights - Solo Clientes que Abandonaron



Insights - Solo Clientes que Abandonaron

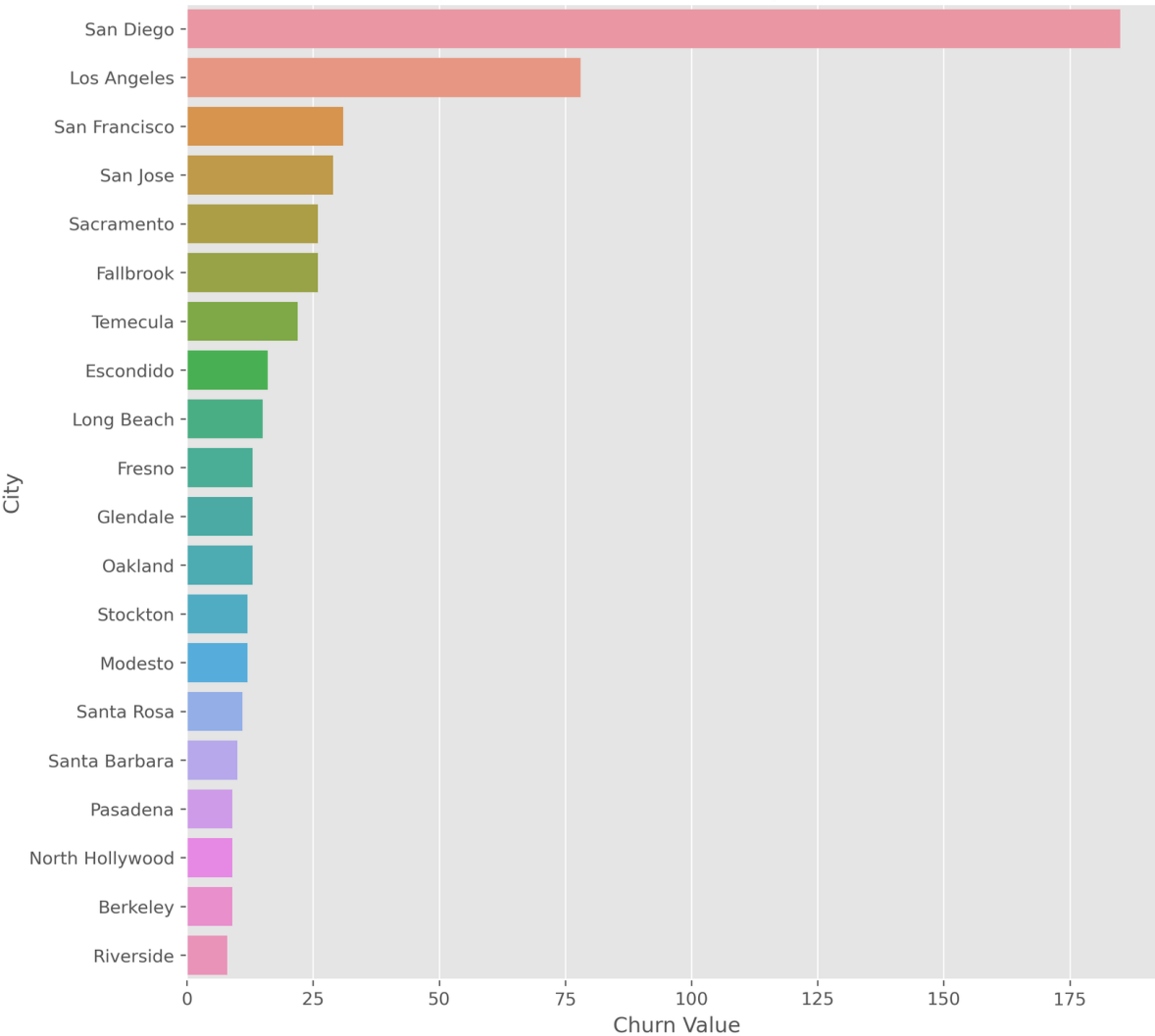


Insights - Solo Clientes que Abandonaron

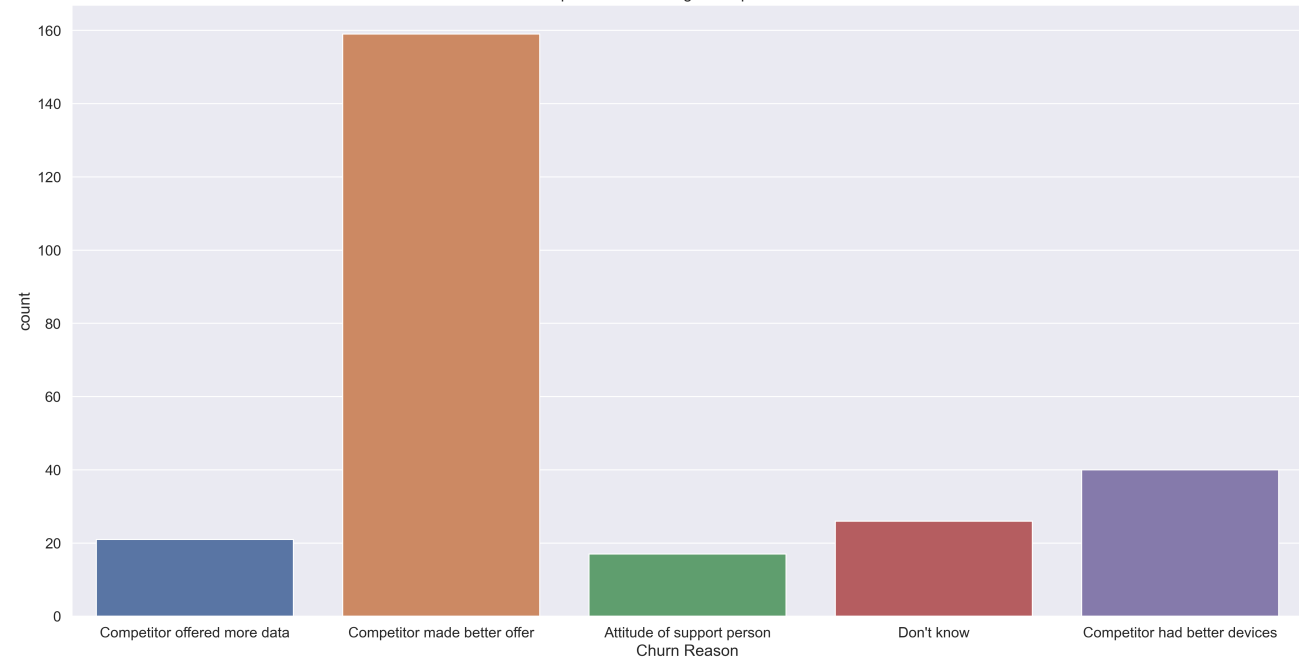


Insights - Solo Clientes que Abandonaron

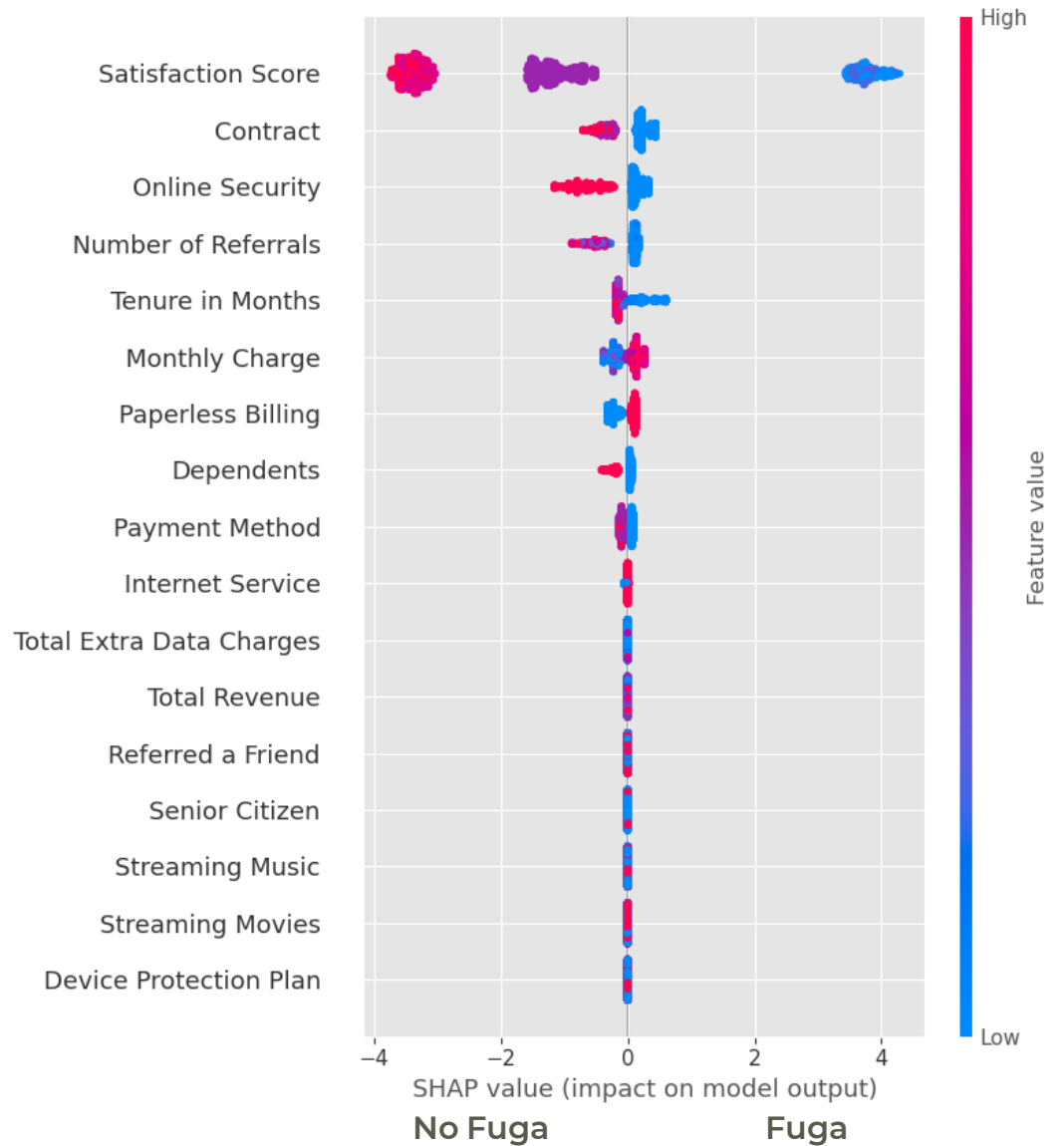
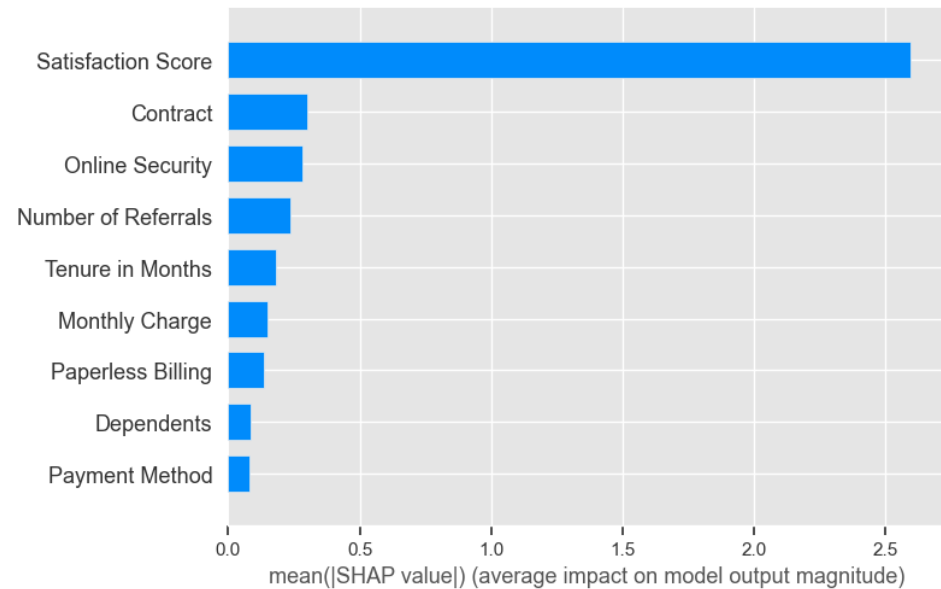
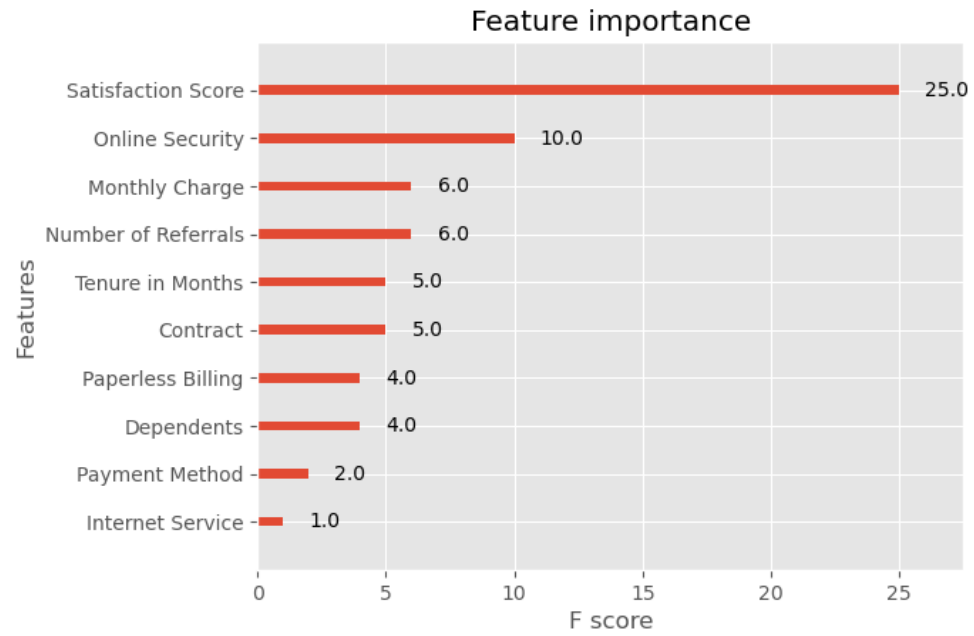
Top 20 Ciudades con más fuga de clientes



Top 5 razones de fuga en top 5 ciudades



Insights - Modelo



Recomendaciones

- Revisar el servicio de Fibra óptica, puesto que es el más solicitado, pero el que presenta más quejas/insatisfacción
- La principal razón de fuga es la competencia. Realizar estudio de mercado, qué ofrece la competencia y qué puede estar haciendo mejor que nosotros
- Establecer paquetes más llamativos y con más beneficios para más contratación anual o de dos años
 - Incluir servicios adicionales como seguridad online, backup online, servicio de protección de planes y asistencia premium
 - Establecer diferenciación con la competencia
- Gran oportunidad de ofrecer ofertas a clientes actuales (la mayoría no cuenta con ninguna)
- Establecer facilidades de pago para clientes que no cuentan con facturación electrónica y/o que pagan con retiro bancario

Recomendaciones

- Establecer encuestas y preguntas a los clientes no solo al momento de la fuga, sino periódicamente para monitorear qué se está haciendo bien
- Preguntar satisfacción de cada tipo de servicio y/o servicio adicional contratado
- Revisar funcionamiento del servicio en San Diego y Los Angeles para determinar si las razones de la fuga de clientes se deben a factores locales o mal funcionamiento de servicios, especialmente en esas zonas
- Utilizar la aplicación del modelo predictivo para tomar acción en clientes que tengan más riesgo de fuga
 - Tener en consideración con los clientes con la variable CLTV (Customer Lifetime Value) con mayor valor
 - Si el cliente con un CLTV alto tiene riesgo de fuga, ofrecer descuentos, mejoras de servicio o estrategias de retención de clientes más agresivas