Case Study: Predict Customer Churn

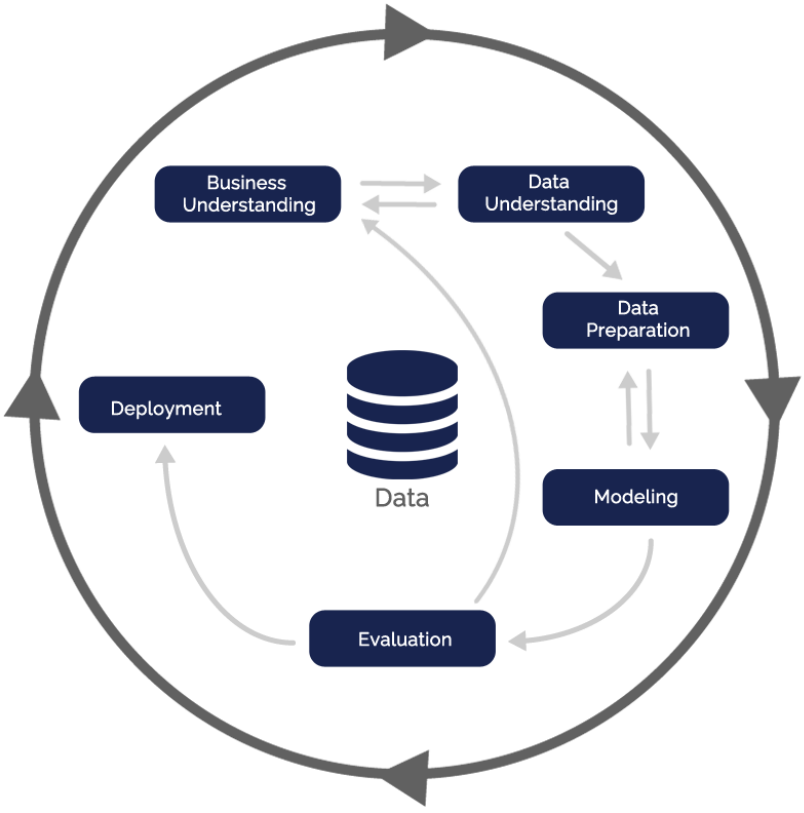
Rikki Mae Martinez

Jennifer Marquez

Patrick Cosmod

5/18/2022

We will see how CRISP-DM is applied in predicting customer churn for SMARTER, a mobile telecommunication provider operating in the North American market.



Note. A Screenshot. CRISP-DM process. Retrieved from https://analyticsindiamag.com/crisp-dm-data-science-project/

During a data science project, Cross Industry Standard Process for Data Mining (CRISP-DM) is used as a process model that represents the foundation for the data science process. The six sequential phases are the following:

1 Business Understanding

1.1 Background

SMARTER Mobile Telecommunication is interested in the reason why customers choose to churn. To study whether churn amounts vary by revenue level, and how often dropped calls occurred in those that churned. Also, a comparison of churning customers and non-churners.

1.2 Executive Summary

Objective:

- Understand why customer is leaving (‘churning’) the company and build a predictive model to analyze who most likely leave in the future.
- Characteristics of those customers who churned and who did not churn.
- Whether customers’ month in service has any impact on churn.
- Whether churn amount was different at certain revenue levels.
- Whether those that churned that had higher levels of dropped or blocked calls.

Dataset Used:

- SMARTER’s Customer Churn Dataset

Tool Used:

- Jupyter Notebook (Anaconda3)

Determine the Best Model

- Decision Tree – best model suited for the dataset.
- Logistic Regression
- K-Nearest Neighbors
- Support Vector Machine
- Naïve Bayes

2 Data Understanding

SMARTER mobile telecommunication has requested us for assistance as they seek to determine which data model to use. An analysis of data was conducted using a classification and regression model. A classification model predicts discrete class labels from observed data. Similarly, regression methods predict continuous quantities. Classification and regression data models are commonly used by many industries to predict the future based on historical data. Our role will be to build a model to predict customer responses for SMARTER mobile telecommunications.

2.1 Data Exploration

- There are in total of **58 variables** (columns) and **50,000 rows** in the dataset
- **Churn** is the target column: 0 or 1
- The dataset is based on the activity of the subscribers' from last month

	customer	revenue	dropvce	blckvce	unansvce	custcare	threeway	mourec	outcalls	incalls	...	retcalls	retaccpt	newcelly	newcelln	refer	income	mc
0	1032537	37.410	15	0	49	7	0	59	47	1	...	0	0	0	0	0	7	
1	1032542	33.395	1	0	0	0	0	27	12	0	...	0	0	0	0	0	0	
2	1032546	32.780	1	0	2	0	0	0	0	0	...	0	0	1	0	0	9	
3	1032549	28.075	1	0	8	6	0	1	6	1	...	0	0	0	0	0	9	
4	1032551	81.555	2	2	16	1	0	97	16	15	...	0	0	0	1	0	8	

5 rows × 58 columns

```
Index(['CUSTOMER', 'REVENUE', 'DROPVCE', 'BLCKVCE', 'UNANSVCE', 'CUSTCARE',
       'THREEWAY', 'MOUREC', 'OUTCALLS', 'INCALLS', 'PEAKVCE', 'OPEAKVCE',
       'DROPBLK', 'CALLFWDV', 'CALLWAIT', 'MONTHS', 'UNIQSUBS', 'ACTVSUBS',
       'PHONES', 'MODELS', 'EQPDAYS', 'AGE', 'CHILDREN', 'CREDITA', 'CREDITAA',
       'PRIZMRUR', 'PRIZMUB', 'PRIZMTWN', 'REFURB', 'WEBCAP', 'TRUCK', 'RV',
       'OCCPROF', 'OCCCLER', 'OCCCRFT', 'OCCSTUD', 'OCCHMKR', 'OCCRET',
       'OCCSELF', 'OWNRENT', 'MARRYUN', 'MARRYYES', 'MAILORD', 'MAILRES',
       'MAILFLAG', 'TRAVEL', 'PCOWN', 'CREDITCD', 'RETCALLS', 'RETACCPT',
       'NEWCELLY', 'NEWCELLN', 'REFER', 'INCOME', 'MCYCLE', 'SETPRC',
       'RETCALL', 'CHURN'],
      dtype='object')
```

```
1  #display dimention of df
2  df.shape
```

```
(49658, 58)
```

- Total Number of Null Values for the entire dataset is **144**
- All Columns win NaN values are **revenue**, **phones**, **models**, and **eqpdays**
- The count of zeros in column age is around **15753**

We need to check the quality of the data. By doing this, we can check whether any data need to be cleansed or to be transformed. For instance, in this dataset, the total number of Null values for the entire dataset is 144. Those columns are revenue, phones, models, and eqpdays or the number of days of the current equipment. And then the count of zeros in column age is around 15753.

	revenue	phones	models	eqpdays
0	37.410	2.0	2.0	299.0
1	33.395	1.0	1.0	733.0
2	32.780	1.0	1.0	564.0
3	28.075	1.0	1.0	626.0
4	81.555	2.0	2.0	666.0
...	...	...	...	...
49653	130.285	2.0	2.0	489.0
49654	NaN	1.0	1.0	773.0
49655	NaN	2.0	2.0	526.0
49656	NaN	1.0	1.0	773.0
49657	NaN	3.0	2.0	378.0

49658 rows × 4 columns

	age
0	23
1	32
2	56
3	0
4	20
...	...
49653	45
49654	40
49655	0
49656	43
49657	0

49658 rows × 1 columns

```
1 (df['age'] == 0).sum()
```

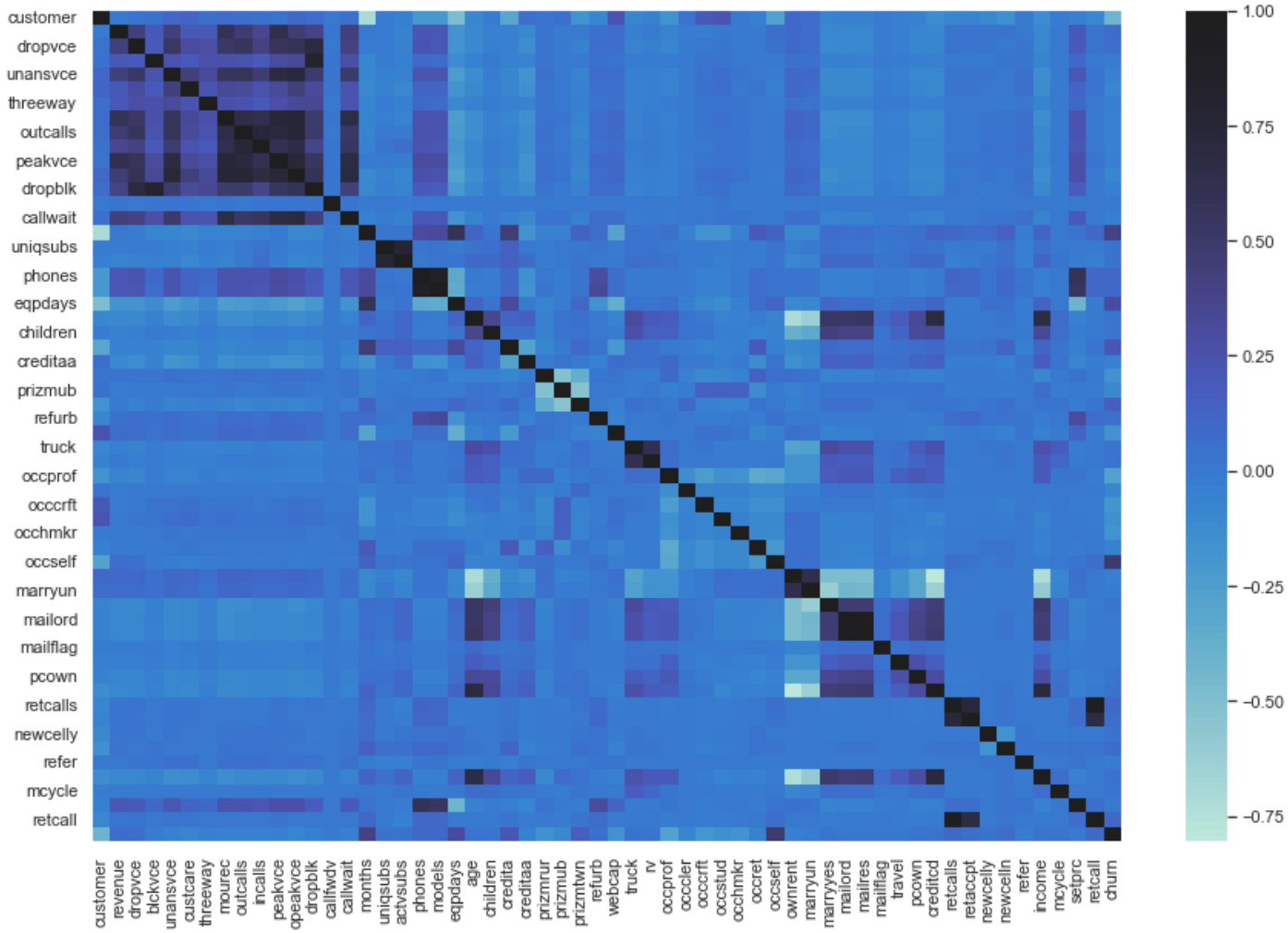
```
15753
```

Attributes info:

1. revenue – revenue from previous month
2. phones – number of handsets issued
3. models – number of models issued
4. eqpdays – number of days of the current equipment

Correlations between different features

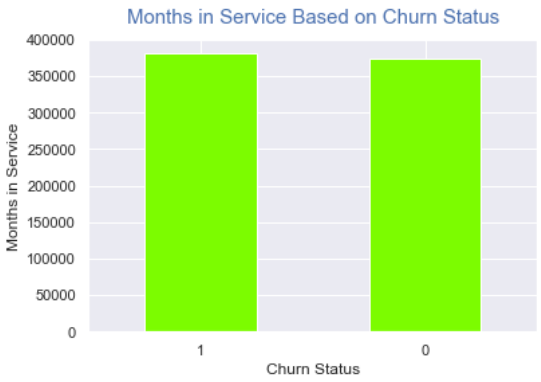
- Based on the heatmap, there are some highly correlated features.
- This may lead to unreliable prediction.

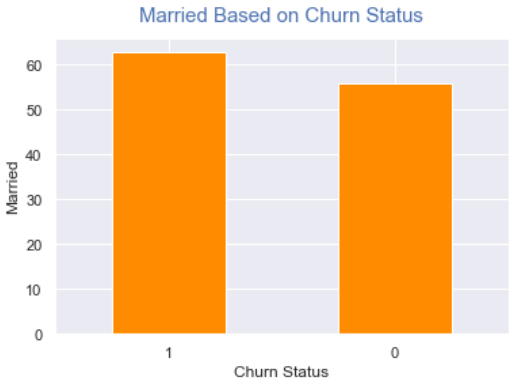


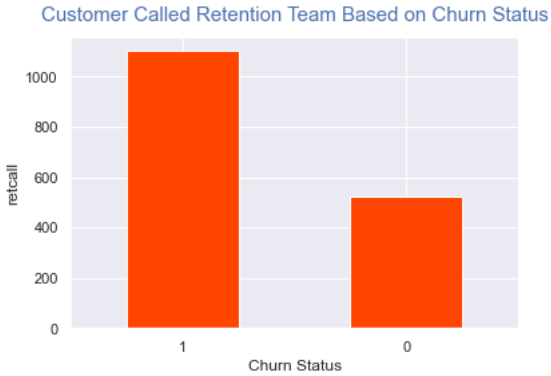
Analyze of churn based on count of different features

- As the customer turns old in service, they are more likely to churn.
- Customers that are married are more likely to churn.
- It can be observed that most of the customers who called retention team churned.

The first graph is about months in service based on churn status, and as the customer turns old in service, they are more likely to churn. In the second graph, married based on churn status, married customers are more likely to churn. And then the last graph, the customer called retention team based on churn status. It can be observed that most of the customers who called the retention team churned. So all of this shows us that we need to cleanse and prepare the data before we do any analysis to capture all relevant information.







Attributes info:

1. months – months in service
2. marryes – married customer
3. retcall – customer has made call to retention team

3 Data Preparation

Some of the following are used for data preparation:

Select data: In order to visualize data more effectively, we choose some of the most relevant characteristics that will be used in data analysis.

Clean data: The dataset has a lot of missing values in fields such as *revenue, phones, models, and eqpdays.* By data imputation, we are able to fill in the inconsistent data with their corresponding mean value. Moreover, the *age* field contains many zero values, so we made zero values to null and filled it with mean values again by imputed data.

Storing data: The dataset contains a target variable, so classification follows the supervised learning approach. It is imperative to first determine which target variable is the model by separating the target variable from other variables and storing them as the response variable and the other field as predicted variables.

3.1 Cleaning

	revenue	phones	models	eqpdays
0	37.410	2.0	2.0	299.0
1	33.395	1.0	1.0	733.0
2	32.780	1.0	1.0	564.0
3	28.075	1.0	1.0	626.0
4	81.555	2.0	2.0	666.0
...	...	...	...	...
49653	130.285	2.0	2.0	489.0
49654	NaN	1.0	1.0	773.0
49655	NaN	2.0	2.0	526.0
49656	NaN	1.0	1.0	773.0
49657	NaN	3.0	2.0	378.0

49658 rows × 4 columns

* All columns with NaN values

* Total number of nulls values in the dataset: **144**

- Fill the missing values in revenue, phones, models, and eqpdays columns with mean values.

```
          revenue  phones  models  eqpdays
0          37.410     2.0     2.0    299.0
1          45.790     2.0     2.0    182.0
2         104.735     7.0     2.0      8.0
3          94.575     3.0     2.0    367.0
4          77.160     3.0     3.0    148.0
...           ...     ...     ...      ...
49653     266.485     4.0     2.0     55.0
49654      72.490     1.0     1.0    251.0
49655      73.775     1.0     1.0    298.0
49656     214.405     2.0     2.0    160.0
49657      32.565     1.0     1.0    267.0

[49658 rows x 4 columns]
```

Attributes info:

1. revenue – revenue from previous month
2. phones – number of handsets issued
3. models – number of models issued
4. eqpdays – number of days of the current equipment

- Data Normalization
- Remove any duplicate rows (if any)
- Drop highly correlated features

```
#Drop highly correlated feature
threshold = 0.9


columns = np.full((df_corr.shape[0],), True, dtype=bool)
for i in range(df_corr.shape[0]):
    for j in range(i+1, df_corr.shape[0]):
        if df_corr.iloc[i,j] >= threshold:
            if columns[j]:
                columns[j] = False
selected_columns = df.columns[columns]
selected_columns
df= df[selected_columns]
```

	customer	revenue	dropvce	blckvce	unansvce	custcare	threeway	mourec	outcalls	incalls	...	retcalls	retaccpt	newcelly	newce
customer	1.000000	-0.011811	0.037092	0.043275	0.099607	0.125394	0.060494	0.074269	0.063093	0.053358	...	-0.084582	-0.060063	-0.100446	-0.1500
revenue	-0.011811	1.000000	0.431733	0.210108	0.430390	0.199410	0.201623	0.577818	0.472400	0.361918	...	0.025942	0.023294	0.014589	0.0351
dropvce	0.037092	0.431733	1.000000	0.184118	0.527372	0.296109	0.269212	0.509229	0.555810	0.398118	...	0.028599	0.020122	0.015022	0.0100
blckvce	0.043275	0.210108	0.184118	1.000000	0.251659	0.192692	0.257434	0.262401	0.239248	0.188991	...	0.004355	0.004293	-0.012232	0.0147
unansvce	0.099607	0.430390	0.527372	0.251659	1.000000	0.427363	0.312784	0.552191	0.575355	0.471125	...	0.022658	0.019312	0.018297	0.0125

5 rows × 58 columns

	customer	revenue	dropvce	blckvce	unansvce	custcare	threeway	mourec	outcalls	incalls	...	creditcd	retcalls	retaccpt	newcelly	newcelln	refer	in
0	1032537	37.410	15	0	49	7	0	59	47	1	...	1	0	0	0	0	0	
1	1032542	33.395	1	0	0	0	0	27	12	0	...	0	0	0	0	0	0	
2	1032546	32.780	1	0	2	0	0	0	0	0	...	1	0	0	1	0	0	
3	1032549	28.075	1	0	8	6	0	1	6	1	...	1	0	0	0	0	0	
4	1032551	81.555	2	2	16	1	0	97	16	15	...	1	0	0	0	1	0	

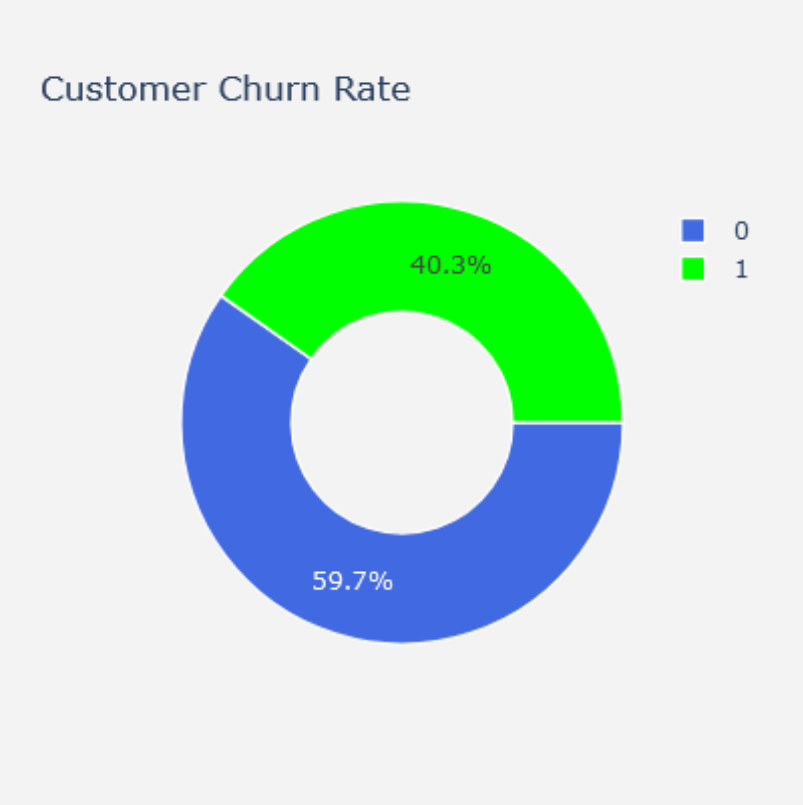
5 rows × 56 columns

4 Modeling

After the data has been pre-processed, particular modeling techniques are chosen, and data can be modeled. To determine the appropriate binary classification model to utilize, such as Logistic Regression, k-Nearest Neighbors, Decision Trees, Support Vector Machines, and Naive Bayes. We split the data into training and test groups to create the design. This approach is used to evaluate and test the quality of various models before interpreting the results.

4. 1 Data Visualization

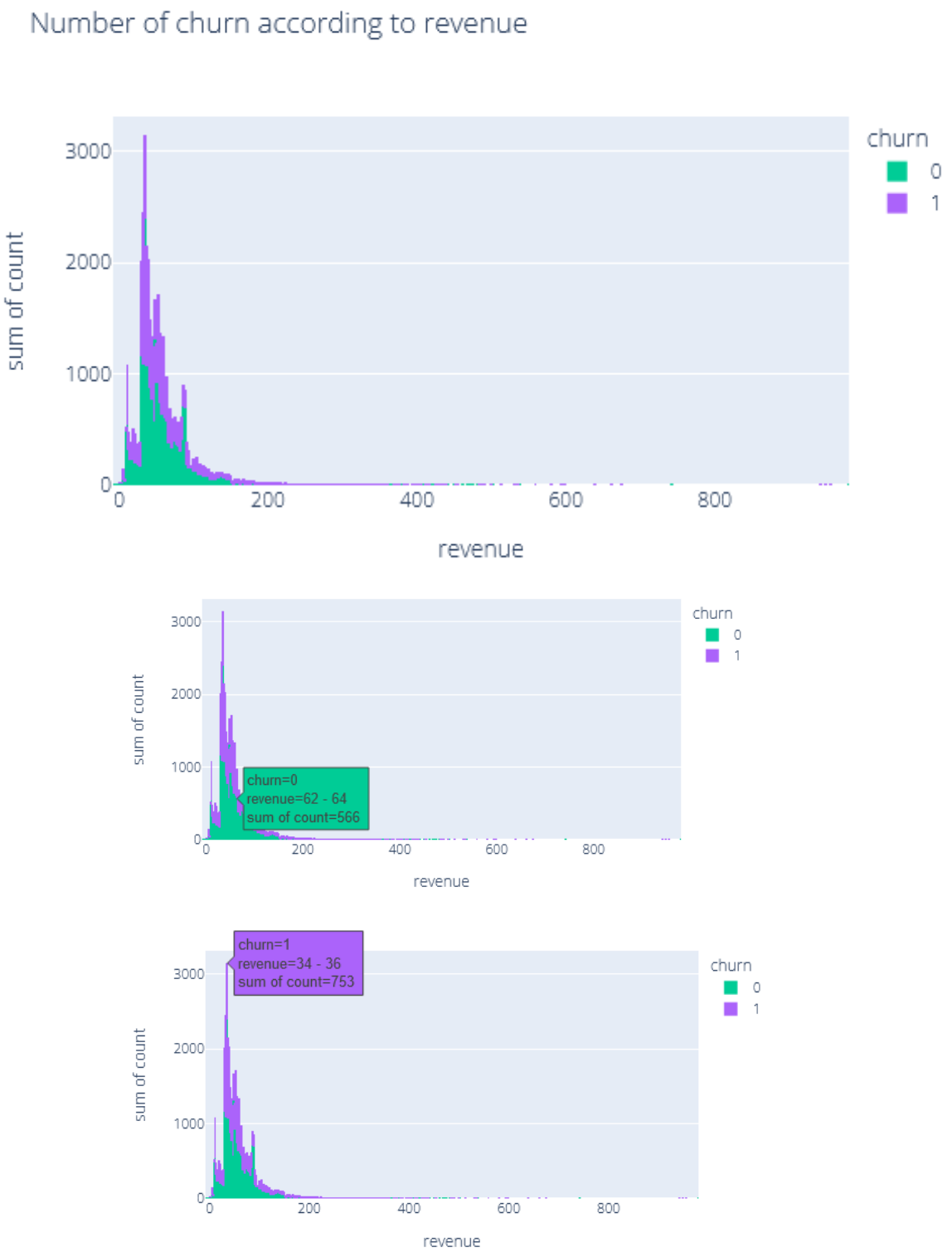
After careful and rigorous data cleansing, we acquire our final data that will be used for analysis and modelling.



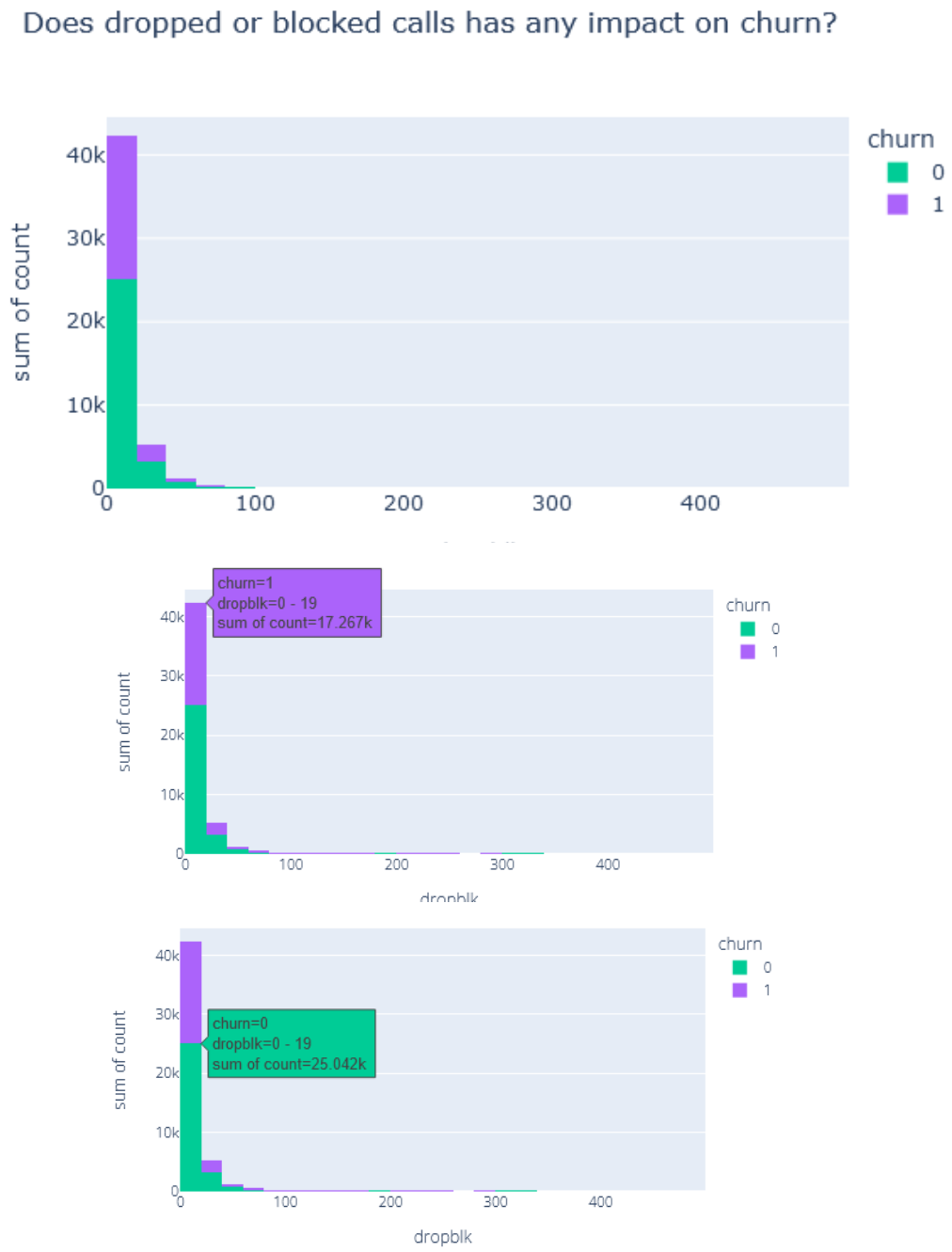
VISUALIZING THE DATA FIRST WITH RESPECT TO CHURN: 1 (Churned), 0 (No

---

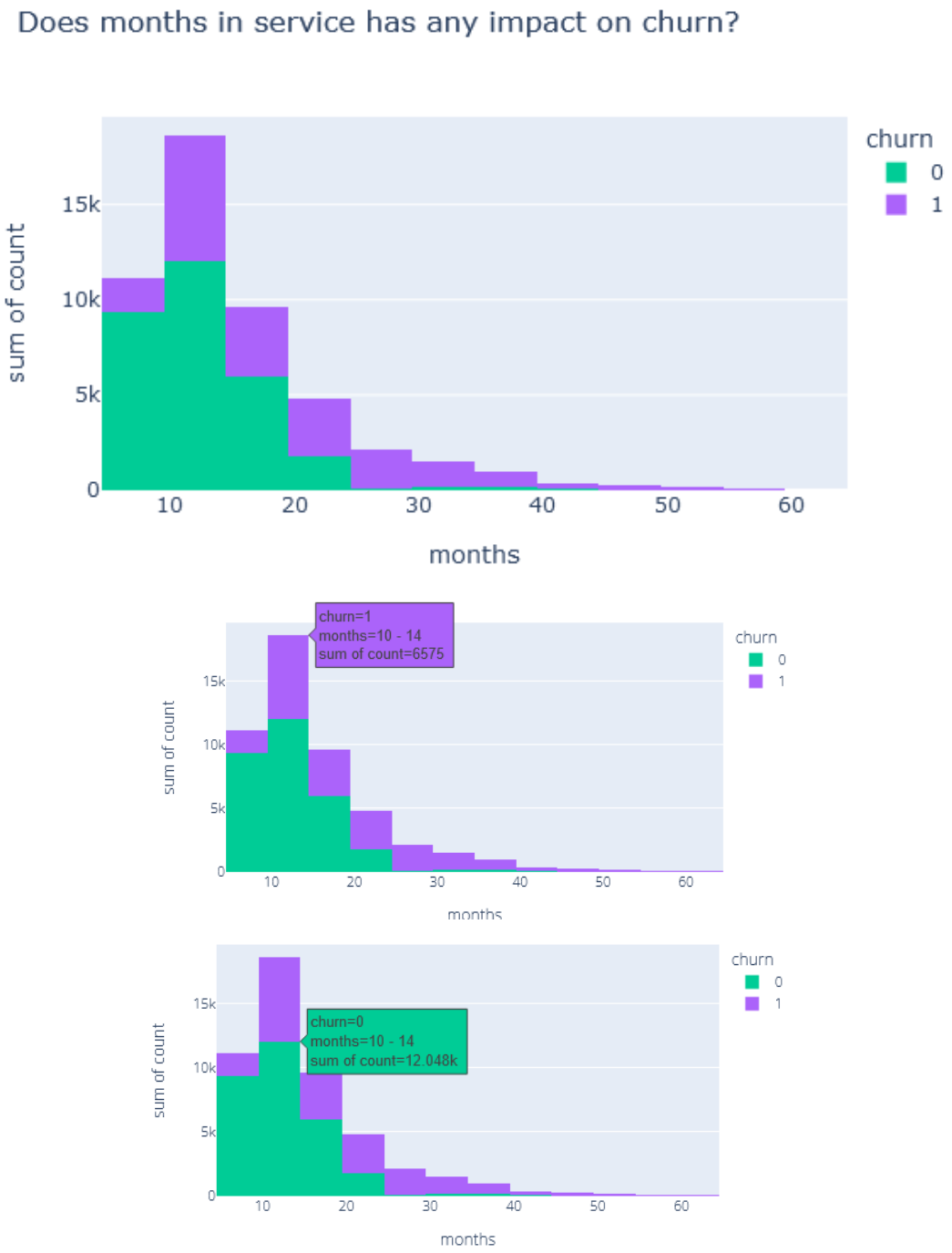
Observation: ‘churn’ column tells us about the number of Customers who left within the last month. Around 40.3% of customers left the platform within the last month.



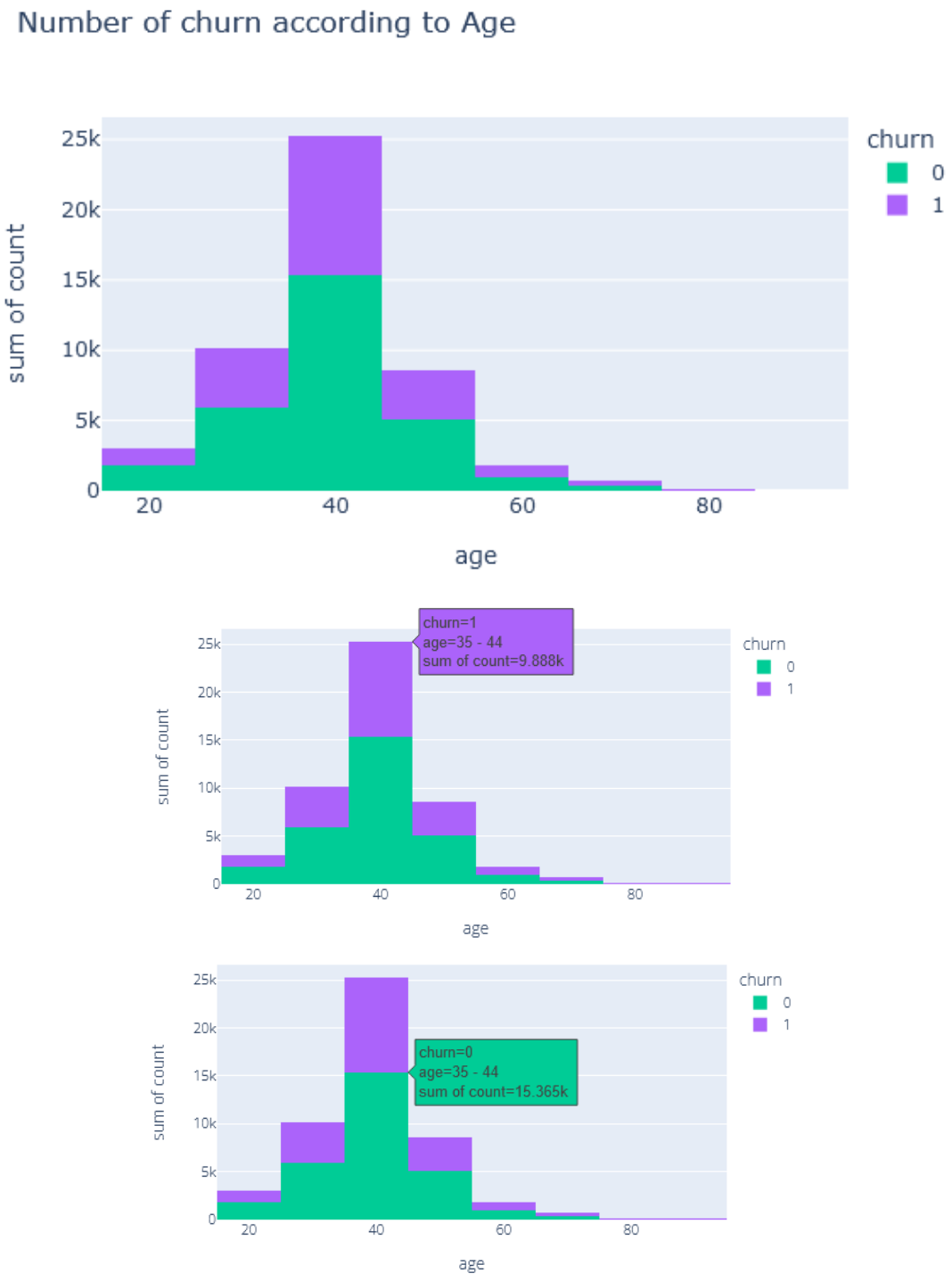
Observation: Churn amount was different at certain revenue levels. However, revenue feature is moderately correlated with customer churn.



Observation: During 0-19 customer's number of dropped or blocked calls, we can see maximum churning. We can see that customers who does not churn had higher levels of dropped or blocked calls.



Observation: during 10-14 of months the customer had service, we can see maximum churning. As the customer turns old, they might get habituated to discontinue using the same telecom service.

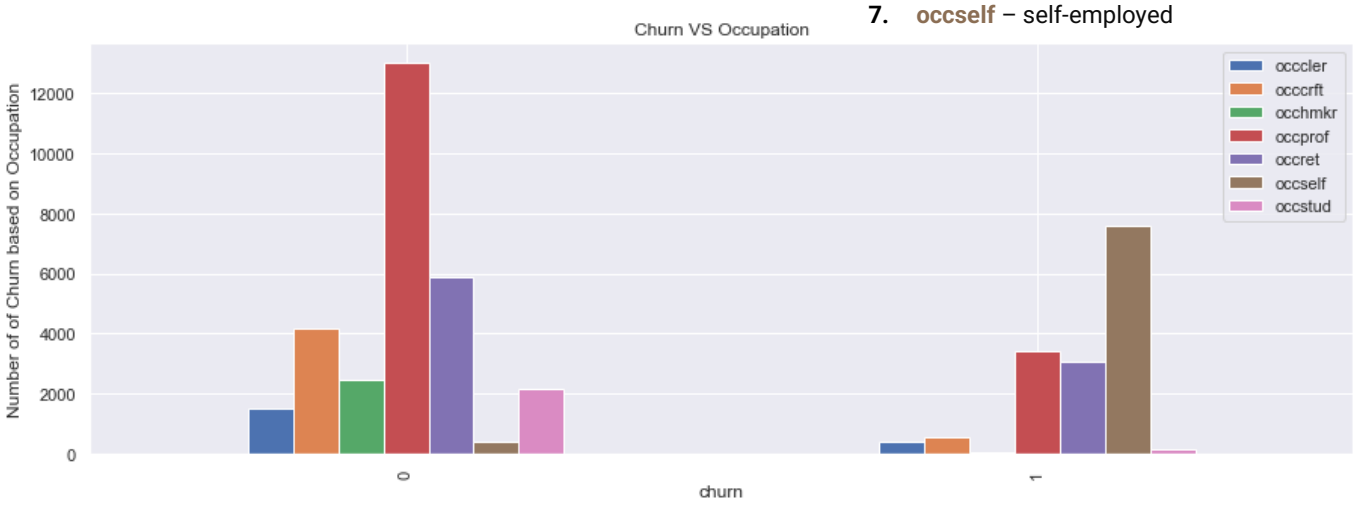


Observation: customer age around 35 - 44, we can see maximum churning. As the customer turns old , they might get habituated to discontinue using the same telecom service.

churn	occcler	occcrft	occhmkr	occprof	occret	occself	occstud
0	1523	4186	2469	12986	5895	431	2168
1	406	563	69	3403	3091	7606	152

Attributes info:

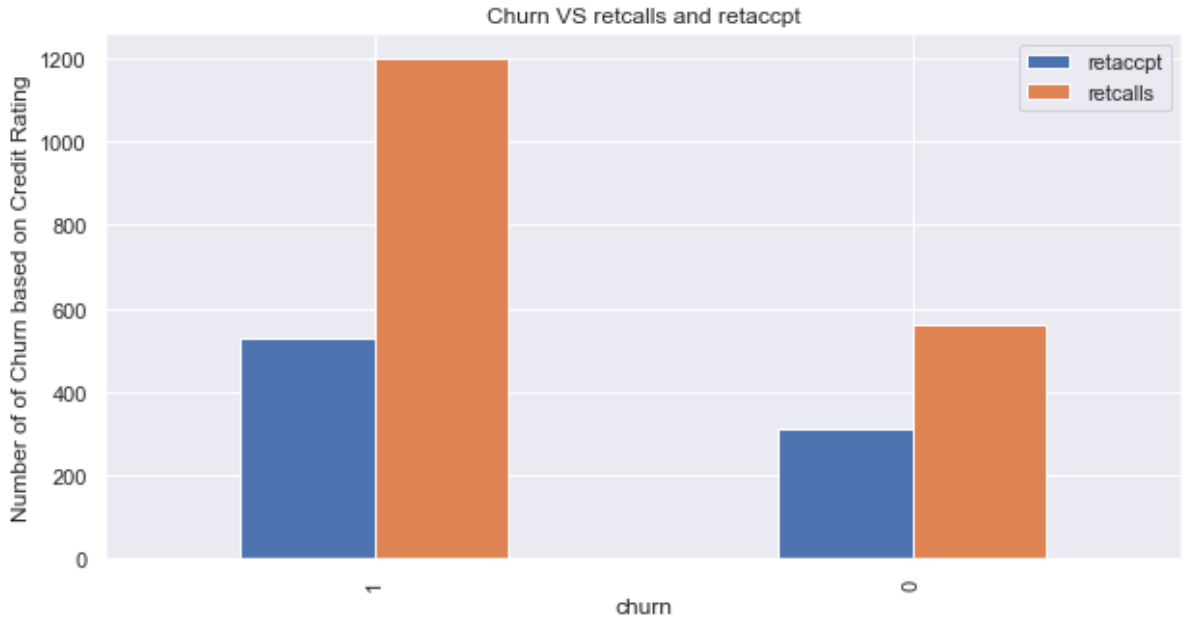
1. **occprof** – professional
2. **occcler** – clerical
3. **occcrft** – crafts
4. **occstud** – student
5. **occhmkr** – homemaker
6. **occret** – retired
7. **occself** – self-employed



churn	retaccpt	retcalls
1	530	1199
0	309	561

Attributes info:

1. **retaccpt** - Number of previous retention offers accepted
2. **retcalls** - Number of calls previously made to retention team

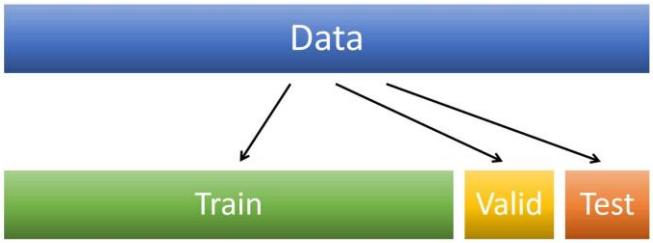


4. 2 Building the Model

- Target
  - Made the “churn” to be target
- Store independent variables to x
  - Obtain overall understanding of data
- Perform test and train split
  - Train + Validation data – **70%**
  - Test data – **30%**

	churn
0	0
1	0
2	0
3	0
4	0
...	...
49653	1
49654	1
49655	1
49656	1
49657	1

49658 rows × 1 c



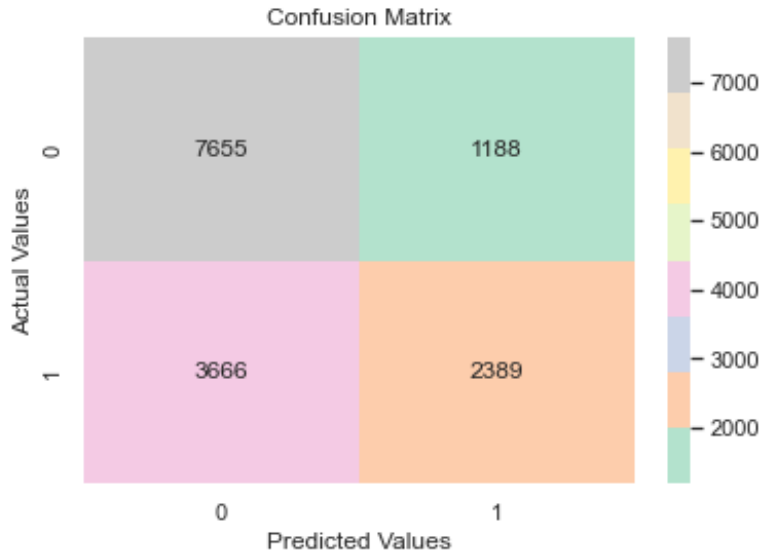

```
1 X.columns
```

```
Index(['revenue', 'dropvce', 'blckvce', 'unansvce', 'custcare', 'threeway',
       'mourec', 'outcalls', 'incalls', 'peakvce', 'opeakvce', 'dropblk',
       'callfwdv', 'callwait', 'uniqsubs', 'actvsubs', 'eqpdays', 'credita',
       'creditaa', 'prizmrur', 'prizmub', 'prizmtwn', 'refurb', 'occprof',
       'occcler', 'occcrft', 'occstud', 'occhmkr', 'occret', 'ownrent',
       'marryun', 'marryyes', 'mailord', 'mailflag', 'creditcd', 'retcalls',
       'retaccpt', 'newcelly', 'newcelln', 'refer', 'income', 'mcycle',
       'setprc'],
      dtype='object')
```

4. 3 Model Evaluation

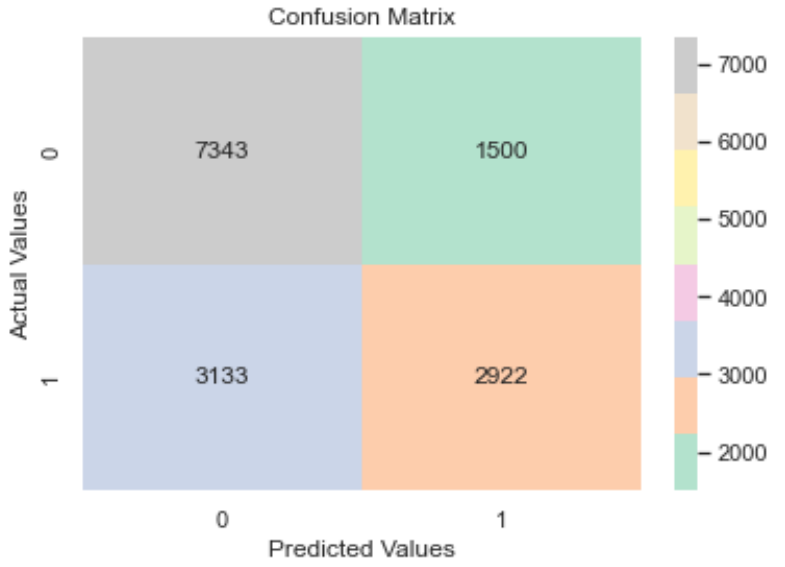
Model 1: Logistic Regression

```
Accuracy Score: 67.42
---------------------------------------------
Actual Values
        churn
6418        0
6748        0
6269        0
20533       0
32531       1
...       ...
13969       0
6496        0
9551        0
44837       1
12273       0

[14898 rows x 1 columns]
Predicted Values
 [0 0 0 ... 0 0 0]
---------------------------------------------
Confusion Matrix :
 [[7655 1188]
 [3666 2389]]
```



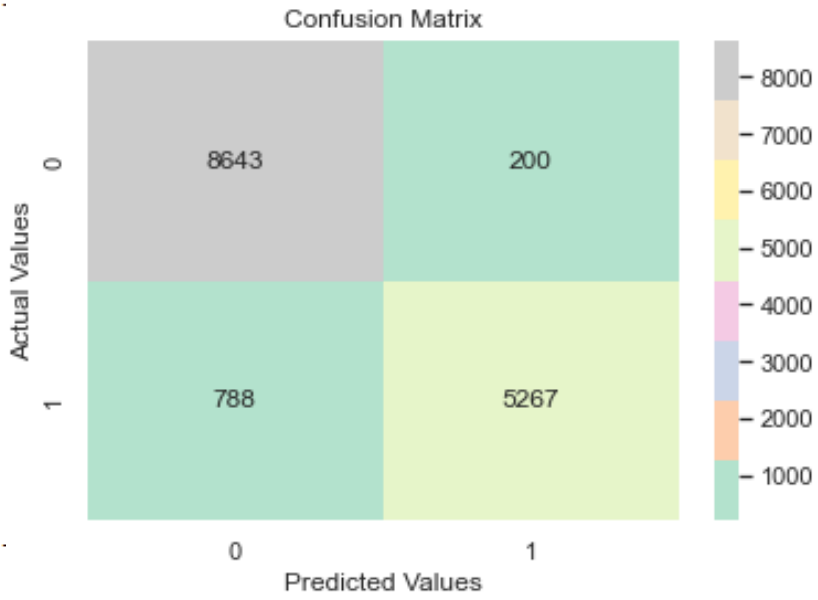
Model 2: K-Nearest Neighbors

```
Accuracy Score: 68.90
-------------------------------------------
Actual Values
        churn
6418        0
6748        0
6269        0
20533       0
32531       1
...       ...
13969       0
6496        0
9551        0
44837       1
12273       0

[14898 rows x 1 columns]
Predicted Values
 [1 0 0 ... 0 0 1]
-------------------------------------------
Confusion Matrix :
 [[7343 1500]
 [3133 2922]]
```



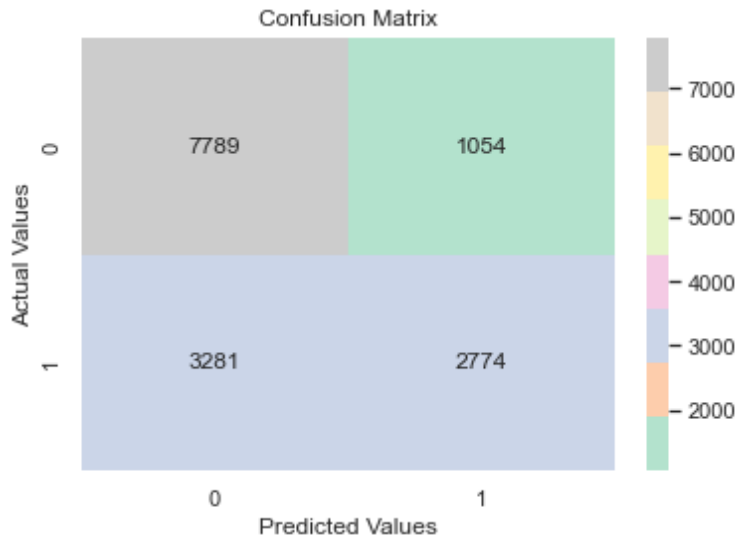
Model 3: Decision Tree

```
Accuracy Score: 93.37
-------------------------------------------
Actual Values
        churn
6418        0
6748        0
6269        0
20533       0
32531       1
...       ...
13969       0
6496        0
9551        0
44837       1
12273       0

[14898 rows x 1 columns]
Predicted Values
 [0 0 0 ... 0 0 0]
-------------------------------------------
Confusion Matrix :
 [[8643  200]
 [ 788 5267]]
```



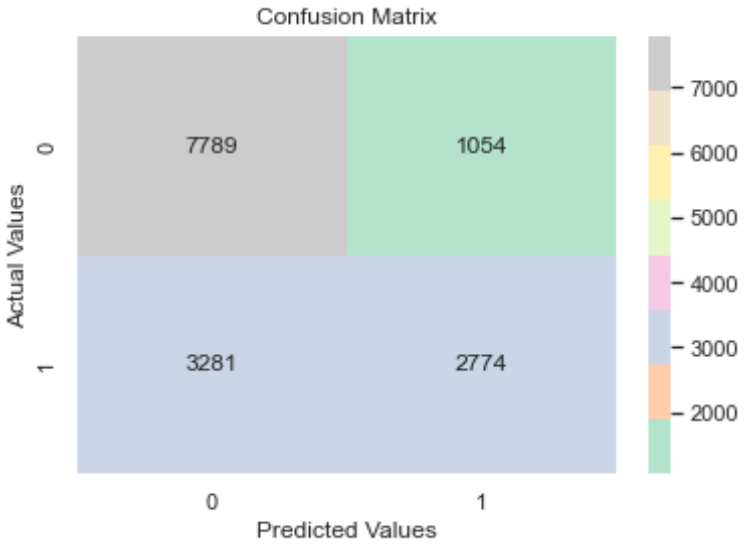
Model 4: Support Vector Machine

```
Accuracy Score: 70.90
---------------------------------------------
Actual Values
        churn
6418        0
6748        0
6269        0
20533       0
32531       1
...       ...
13969       0
6496        0
9551        0
44837       1
12273       0

[14898 rows x 1 columns]
Predicted Values
 [0 0 0 ... 0 0 0]
---------------------------------------------
Confusion Matrix :
 [[7789 1054]
 [3281 2774]]
```



Model 5: Naïve Bayes

```
Accuracy Score: 70.51
---------------------------------------------
Actual Values
        churn
6418        0
6748        0
6269        0
20533       0
32531       1
...       ...
13969       0
6496        0
9551        0
44837       1
12273       0

[14898 rows x 1 columns]
Predicted Values
 [0 0 0 ... 0 0 0]
---------------------------------------------
Confusion Matrix :
 [[6987 1856]
 [2537 3518]]
```



Identify the best model

- From this table, the **Decision Tree Model** is the best model as it has the accuracy score of 93%.

Score	Model
0.933280	Decision Tree
0.709021	Support Vector Machines
0.705128	Naive Bayes
0.677742	K-Nearest Neigbor
0.674184	Logistic Regression

5 Evaluation

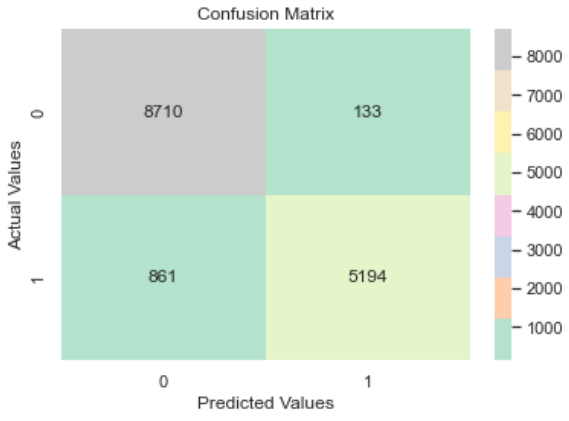
The achieved model was properly examined throughout the evaluation step. The most accurate model of binary classification is the Decision Tree, which has a 93 percent accuracy rating. In addition, the F1 score is 91.27 percent, while the Precision score is 97.50 percent, and the Recall score is 85.78 percent. The confusion matrix is also analyzed and interpreted properly.

5. 1 Approved Model

Decision Tree Model- Accuracy of the Classifier

```
Accuracy score: 93.33
Precision: 97.50
Recall: 85.78
F1 score: 91.27
```

```
Confusion Matrix :
 [[8710  133]
 [ 861 5194]]
```



Interpretation:

TN: eight thousand seven hundred ten (8710) customers did not churn that are correctly predicted.

TP: five thousand one hundred ninety-four (5194) customers churned that are correctly predicted.

FN: eight hundred sixty-one (861) customers churned, but predicted that they will not churn.

FP: one hundred thirty-three (133) customers did not churn but predicted that they will churn.

6 Deployment

6.1 Conclusion

From the analysis, we can conclude:

- Based on the results from the model comparison, **Decision Tree model** performed the best in order to predict customer churn.
- Customers who are **married**, **called retention team**, age **35 – 44,** and **self-employed are** the most important features that have strong association with churn probability of the customer.
- As the customer **turns old in service**, they are more likely to churn.
- Churn amount was different at certain revenue levels. However, revenue feature is

6.2 Recommendations

- During 10 – 14 months the customer had service, we can see maximum churning. As they turn old in service, they might get habituated to discontinuing using the same telecom service. **The sales team may focus on customers who have not made a recharge in the last 10 – 14 months or more and provide them with special talk-time offers to reduce churn.**
- Customers who previously called the retention team are more likely to churn. **Telecom's retention team may need to improve their service to retain customers using the same platform. We need to hire the correct candidates for the retention team or encourage them to learn proactively through seminars and conferences.**
- Customers who are self-employed are more likely to churn. **The sales team may focus on the self-employed customer to provide recommendations about the subscription's benefits and advantages and what sets us distinct from the competitors.**
- Customers who are senior citizens got the lowest rate for no-churn. **Customer support may focus on senior citizens by assisting them with the technical side. We need to increase senior customer engagement by offering one-on-one consultations for struggling customers.**
- Lastly, for overall impact to reduce customer churning. The telecom company may provide email marketing promotional campaigns to keep customers. We can let them know what discounts and options are best for them using email promotional offers.