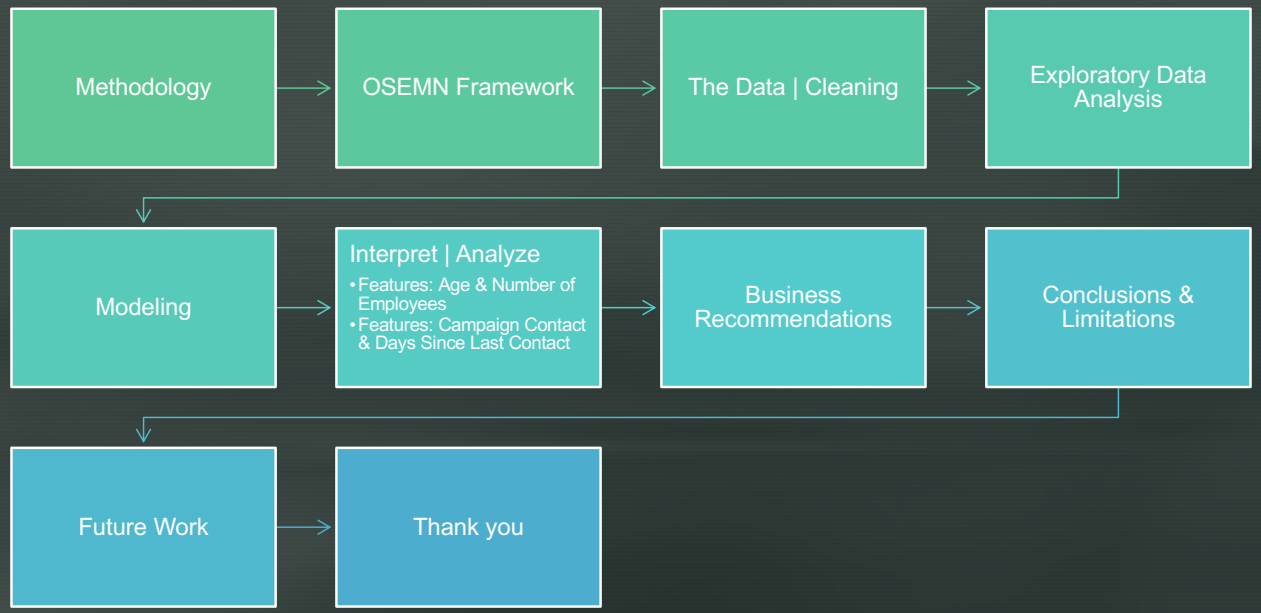
Telecommunications Analysis Customer Churn

Miguel Santana

Flatiron School

Data Science | FT Cohort

Introduction



Methodology

A telecommunications company (SyriaTel) is looking to improve upon customer churn ratios. SyriaTel provided client data to better understand why customers stop doing business with the company.

Our team leveraged the OSEMN framework to analyze the dataset and make key business recommendations .

Data Science Process

LEAD

OBTAIN | SCRUB | EXPLORE | MODEL | INTERPRET

O — Gather data from relevant sources

S — Clean data to formats that machine understands

E — Find significant patterns and trends using statistical methods

M — Construct models to predict and forecast

N — Put the results into good use

OSEMN Framework

Originally by Hilary Mason and Chris Wiggins

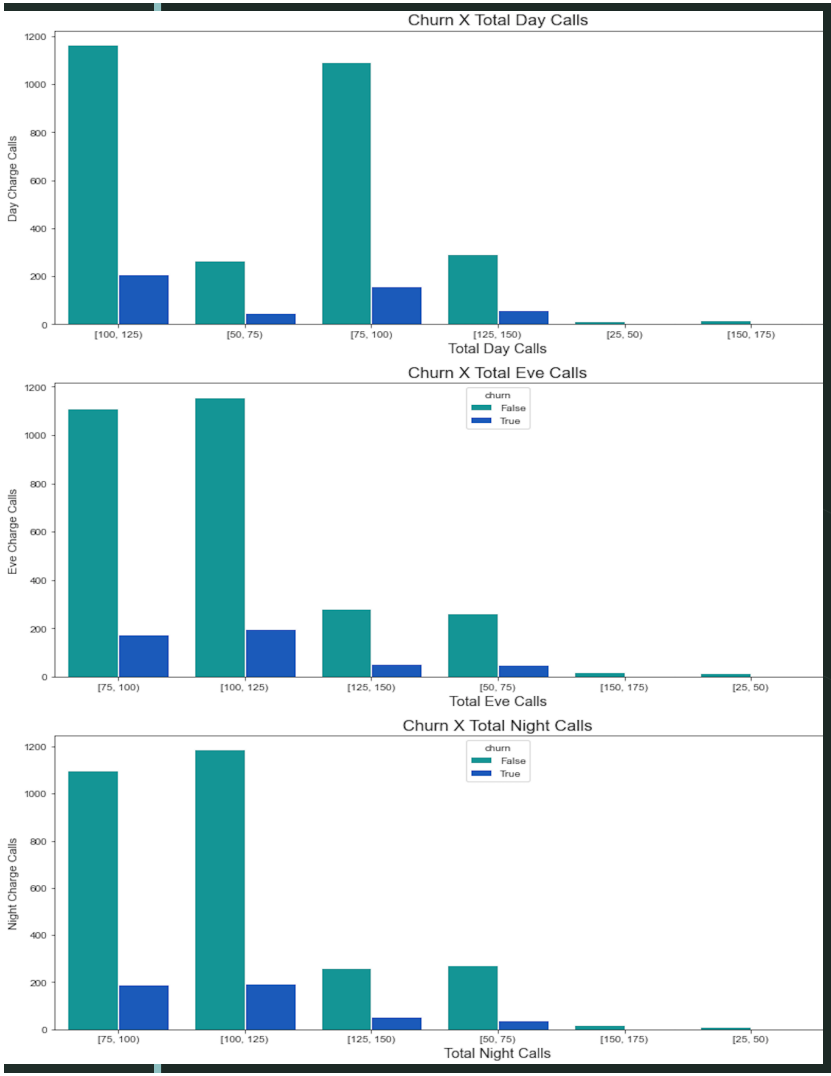
The Data | Cleaning

The SyriaTel dataset is available on Kaggle's website. The dataset can be found:

- https://www.kaggle.com/becksddf/churn-in-telecoms-dataset

Cleaning the data

- Addressing null and abnormal values
- Feature Selection
- Maintaining as many of the dataset values as possible



EXPLORATORY DATA ANALYSIS

Most Common Client Features

- Day, Eve & Night Calls:
  - Highest churn occurs when customers make between 75-125 days calls per billing cycle.

Modeling

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
0	Light Gradient Boosting Machine	0.9477	0.9059	0.7707	0.8562	0.8103	0.7801	0.7821	0.4736
1	CatBoost Classifier	0.9472	0.9051	0.7708	0.8547	0.8100	0.7795	0.7812	3.3868
2	Gradient Boosting Classifier	0.9373	0.9028	0.7771	0.7909	0.7832	0.7465	0.7471	0.7334
3	Extreme Gradient Boosting	0.9355	0.9083	0.7709	0.7871	0.7763	0.7387	0.7405	0.2347
4	Random Forest Classifier	0.9093	0.8884	0.5482	0.7778	0.6382	0.5883	0.6028	0.1104
5	Extra Trees Classifier	0.8994	0.8796	0.4616	0.7568	0.5713	0.5182	0.5397	0.1631
6	Decision Tree Classifier	0.8895	0.8212	0.7249	0.6051	0.6577	0.5927	0.5974	0.0313
7	Ada Boost Classifier	0.8565	0.8287	0.5888	0.5092	0.5442	0.4600	0.4628	0.2084
8	Logistic Regression	0.7599	0.7938	0.6936	0.3413	0.4571	0.3253	0.3593	0.0543
9	K Neighbors Classifier	0.7581	0.7857	0.6844	0.3378	0.4518	0.3190	0.3522	0.0075
10	Ridge Classifier	0.7514	0.0000	0.6966	0.3324	0.4496	0.3144	0.3506	0.0065
11	Linear Discriminant Analysis	0.7500	0.7921	0.6936	0.3304	0.4471	0.3112	0.3472	0.0348
12	SVM - Linear Kernel	0.7229	0.0000	0.7372	0.3179	0.4405	0.2968	0.3445	0.0349
13	Naive Bayes	0.5271	0.5964	0.5821	0.1705	0.2634	0.0495	0.0707	0.0036
14	Quadratic Discriminant Analysis	0.5104	0.5825	0.6069	0.1735	0.2656	0.0532	0.0738	0.0093

The top model selected for feature evaluation is Light Gradient Boosting Machine

Interpret | Analyze

- Generated most impactful features
- Top Features Include
  - Customer Service Calls
  - Total Day Minutes
  - Total International Minutes

Business Recommendations

- Churn is highest during the first 3 customer service interactions. SyriaTel should deploy an A-Team of high performers to address incoming service calls from new customers.
- Total day minutes make up the majority of the minutes consumed and therefore relay to the largest percentage of the total cost. SyriaTel can create and market advertisements for consuming minutes during low peak business hours or consider changing their day pricing (adjusting eve, night and intl costs to compensate).
- Churn grows exponentially as consumers go over 7 international minutes. SyriaTel should create an outreach campaign to inform customers when they are going over the 5 international minute threshold. SyriaTel should also consider creating a forgiveness program for first time customers who consumed large amounts of international minutes without knowing the cost (one time only).

Conclusion

The dataset offered various consumer trends and illustrated multiple areas of opportunity. Two areas of opportunity addressed common retail business pitfalls; customer service and high cost products (international minutes being the most expensive). Lastly, day minutes are the most common of the available categories and directly influenced the majority of consumer costs. These and future trends can be used to prepare for future business.

Future Work

Future Work In order to more accurately define the boundaries of our features it is important to understand what customs and cultural influences are tied to this dataset. SyriaTel is not a United States based cell phone carrier and cultural influences may impact the way we perceive the information. In addition, it would be helpful to have more client specific data in order to understand on an individual level which client segments are leaving the company and what features they share.

THANK YOU!

Questions? Miguel Santana | contact: msantana269@gmail.com

Github Repo: https://github.com/msantana269/Module-3-Project

Appendix

Model Validation