Customer Churn

...and how Machine Learning can prevent it.

Agenda

Intro

Data

Analysis

Modeling

Recomm.

Intro

What is customer churn?

- Customer churn happens when customers stop using a company’s products or services
- Churn rate is an important metric because **losing customers** equals **losing revenue**
- Hence, losing customers requires gaining new customers
- Acquiring a new customer could **cost 10 times more** than retaining an existing one
- Thus, companies who prevent churn can build a **competitive advantage** in the market

What should companies do?

- Companies need a **retention strategy** in order to avoid an increase in churn rates
- Churn rates vary by industry and knowing your market is key to reducing churn
- Understanding **potential churn signs** and being **proactive** could be key
- The scope of this project is to identify **patterns** between churned customers in telecom
- Ultimately, see if we can successfully **detect** and **prevent churn** using machine learning

Data

What available data did we have?

Customer Demographics

- Gender
- Age
- Partners
- Dependents
- etc.

Account Information

- Tenure
- Contract type
- Charges
- Payment method
- etc.

Service Information

- Phone service
- Multiple lines
- Internet service
- Tech support
- etc.

Descriptives

7.032
customers

26%
churned customers

32 months
avg. tenure

65€
avg. monthly charges

5
avg. number of services

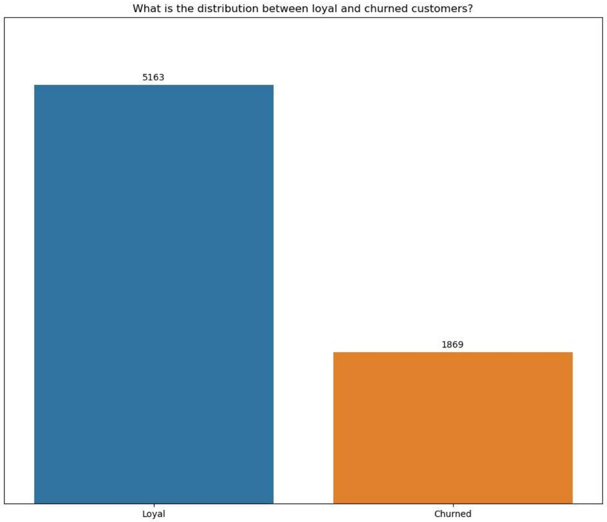
Analysis

Classes

There is an **imbalance** in our dataset between loyal and churned customers.

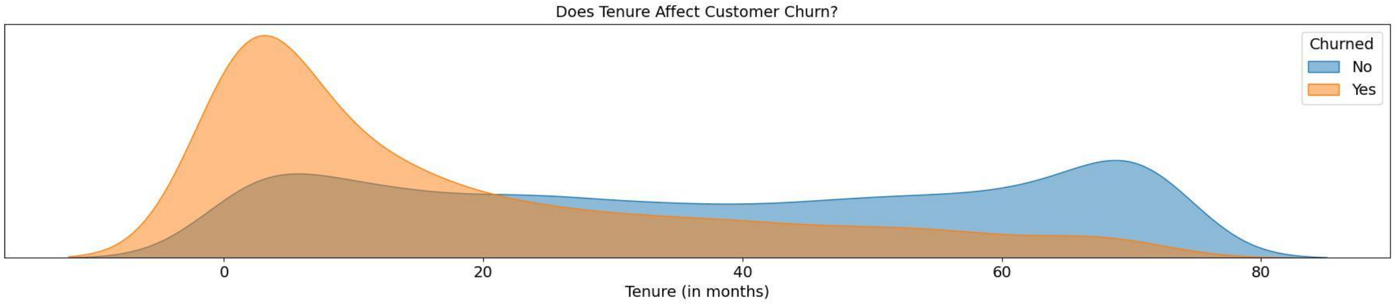
This is something we need to address **before** training our model, otherwise it will introduce **bias** into the results.

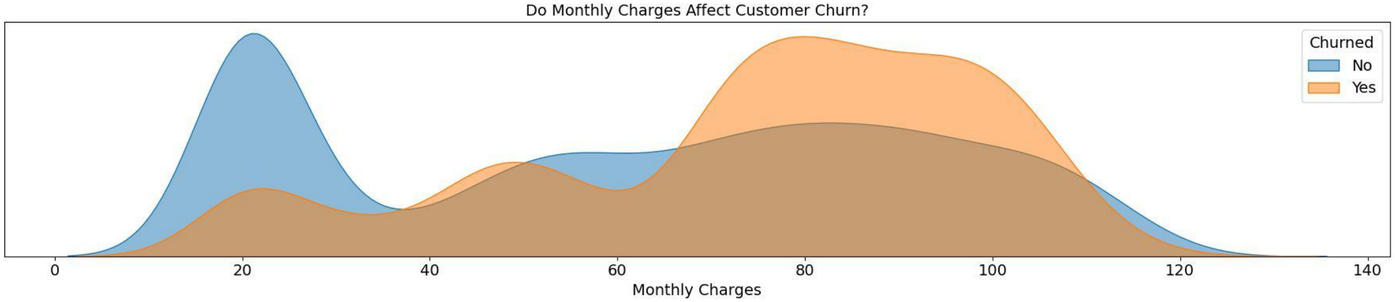
We will explain later how we will try to fix this issue using an **undersampling** technique with **cluster centroids**.



Demographics		Population	% Churned
Gender	F	3483	27 %
	M	3549	26 %
Senior Citizen	N	5890	24 %
	Y	1142	**42 %**
Has Partner	N	3639	33 %
	Y	3393	20 %
Has Dependents	N	4933	**31 %**
	Y	2099	16 %

- Approx. 1 out of 2 **senior citizens** seem to churn
- Also, customers with **dependents** are more likely to churn





- Customers tend to churn more often during the **first few months** of their tenure
- In addition, **higher monthly charges** seem to push customers away

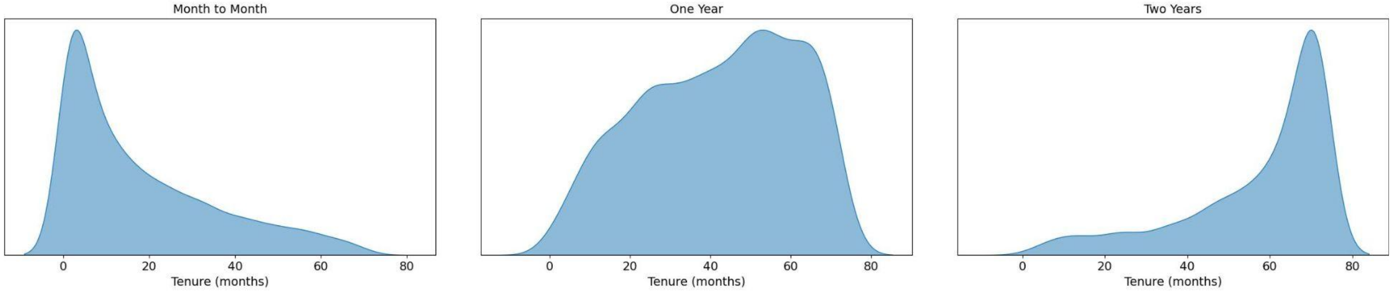
Core Services		Population	% Churned
Has Phone Service	N	680	25 %
	Y	6352	27 %
Has Internet Service	N	1520	**7 %**
	Y	5512	**32 %**
Has DSL	N	4616	31 %
	Y	2416	19 %
Has FIber Optic	N	3936	15 %
	Y	3096	**42 %**

- Customers with **no internet service** does not seem to churn, however...
- Customers with **internet service**, and especially with **fiber optic**, are quite likely to churn

Extra Services		Population	% Churned
Has Online Security	N	5017	**31 %**
	Y	2015	15 %
Has Online Backup	N	4607	29 %
	Y	2425	22 %
Has Tech Support	N	4992	**31 %**
	Y	2040	15 %
Has Streaming TV	N	4329	24 %
	Y	2703	30 %

- Approx. 1 out of 3 customers without **online security** or **tech support** seem to churn
- None of the other services seem to significantly drive customer churning

Contract		Population	% Churned
Duration	Month to Month	3875	**43 %**
	One Year	1472	**11 %**
	Two Years	1685	**3 %**



- Approx. 1 out of 2 customers with **month-to-month** contracts are more likely to churn
- On the other hand, customers with **long-term** contracts seem to be more loyal

Payment Method		**Population**	**% Churned**
Electronic Check	N	4667	17 %
	Y	2365	**45 %**
Mailed Check	N	5428	29 %
	Y	1604	19 %
Bank Transfer	N	5490	29 %
	Y	1524	17 %
Credit Card	N	5511	**30 %**
	Y	1521	15 %

- Customers who pay their bill via **electronic check** tend to churn
- On the contrary, customers who pay by **credit card** are more loyal

Key Findings

- The dataset is highly **imbalanced** and we need to handle this before model training
- **Senior citizens** and customers with **dependents** are more likely to churn
- Customers with **smaller tenure** and **short-term contracts** are more prone to churn
- Moreover, customers with **higher monthly charges** have a higher chance of leaving
- Customers with **internet service**, and especially **fiber optic**, are very prone to churn
- **Electronic check** is the preferred payment method, but those who use it tend to churn

Modeling

Methodology

Data Preprocessing

- Data Cleansing
- One-Hot Encoding

Feature Engineering

- Feature Augmentation
- Feature Transformation
- Feature Scaling

Class Imbalance

- Cluster Centroids

Models & Evaluation

- Cross Validation
- HP Tuning
- Modeling

Preparing the dataset

Step 1

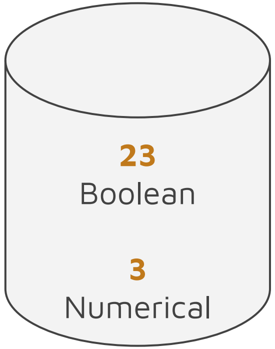
Cleaned and augmented the initial dataset of **20 features** by creating **1 new feature**

Step 2

Converted the remaining **categorical** features into **dummies**

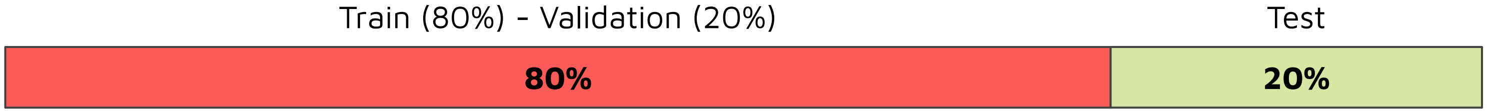
Step 3

Transformed and **scaled** any features with high variance

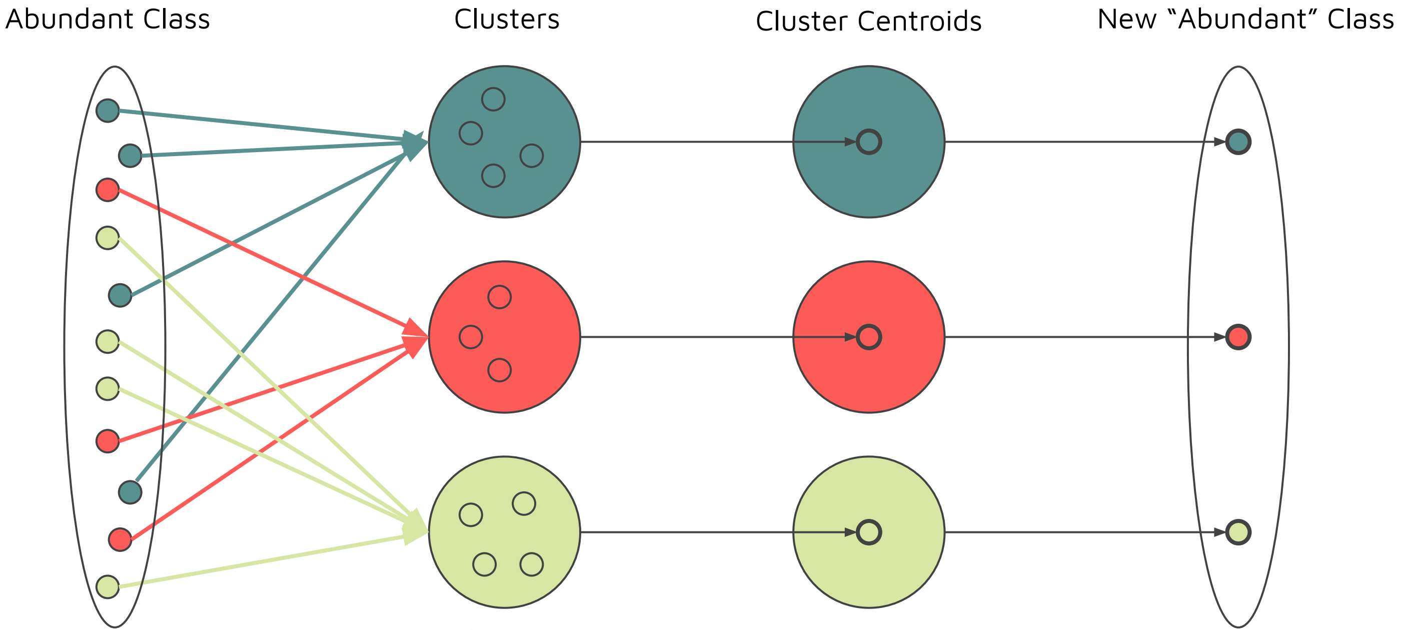


23
Boolean

3
Numerical



Handling class imbalance



New Abundant Class Size = Rare Class Size

** The number of clusters is equal to the number of the data points of the rare class where abundant class > rare class*

Why using cluster centroids?

Pros

- The two classes are now equally balanced, hence no risk of introducing **bias** during model training
- Cluster centroids can be **more representative** data points for the abundant class case

Cons

- After reducing the number of data points within the “abundant” class, the dataset has been **decreased**
- Reducing the size of the dataset can lead to less accurate results due to **loss of information**

Whatever method we use will help in some ways, but hurt in others.
There is no best approach or model for all problems.
It is strongly recommended to try different techniques and models to evaluate what works best.

How did we evaluate the results?

Before we dive into the results of our models, let’s try to understand some of our metrics:

Precision: How many of the predicted customers had actually churned?

Recall: How many of the customers that had actually churned the model predicted right?

F1 Score: The harmonic mean of precision and recall

AUC: Shows how much the model is capable of distinguishing between the two classes

Modeling results

Model	Precision	Recall	F1 Score	AUC	Accuracy
Logistic Regression	0.51	0.80	**0.62**	**0.76**	0.74
SVC	0.47	0.82	0.60	0.74	0.70
RandomForest	0.45	0.82	0.58	0.73	0.69
KNN	0.44	0.85	0.58	0.73	0.67
LightGBM	0.42	**0.89**	0.57	0.73	0.65

** Results sorted by F1 Score*

Confusion matrix (1/2)

Model 1: Logistic Regression

Logistic Regression shows the **highest F1 Score** compared to the rest of the models, which means that it has the **best balance** between **precision** and **recall** metrics.

Also, compared to LightGBM, it has a **lower** number of **False Positive (FP)**, which means that it makes **less mistakes** on predicting a customer as churned, helping the company be more cost-effective in its retention campaigns.

Log. Reg.	**Predicted: 0**	**Predicted: 1**
Actual: 0	741 **TN**	292 **FP**
Actual: 1	75 **FN**	299 **TP**

Comment

Cost-effective as it will hardly label a non churned customer as churned, however it will not “catch” a lot of potentially churned ones.

Confusion matrix (2/2)

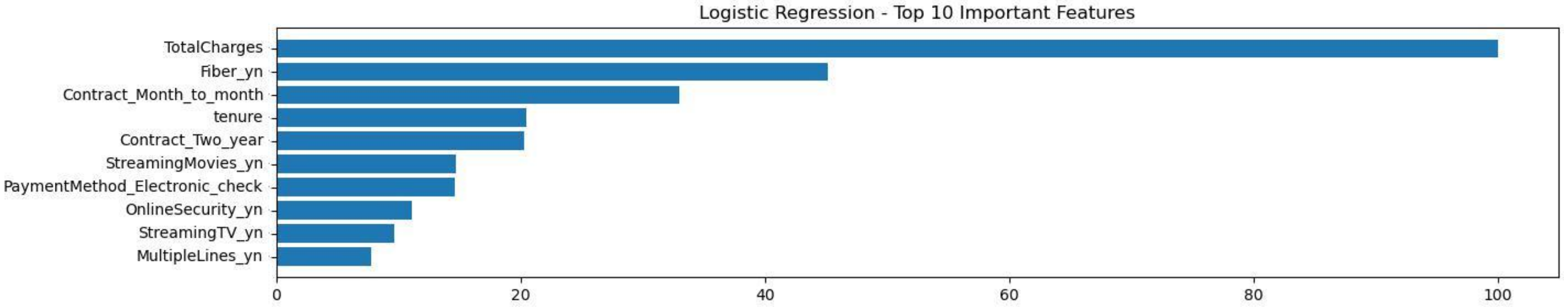
Model 2: LightGBM

LightGBM model leads to a **higher** number of **False Positive (FP)** predictions. In practice, this can be bad because it means that the model includes in his predictions as churned customers, customers who did not churn.

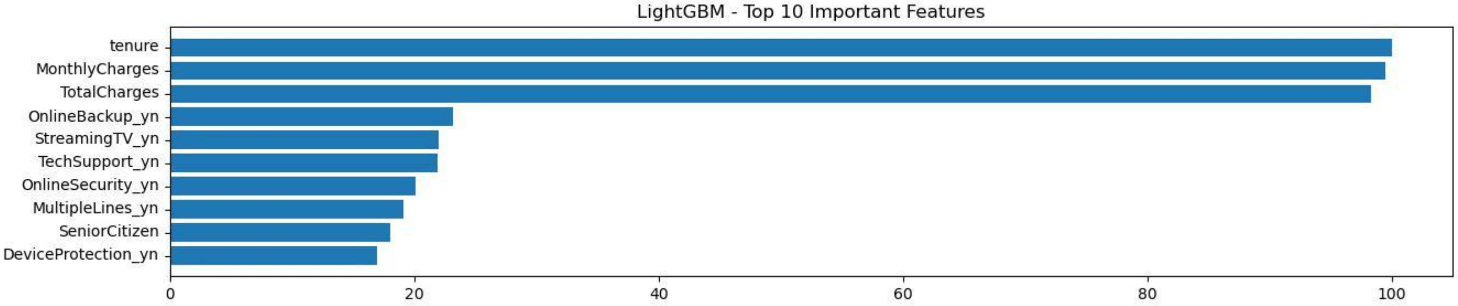
On the other hand, it returns a **higher** number of **True Positive (TP)** predictions, which is good, since it gives a more accurate number of churned customers than the previous model.

LightGBM	Predicted: 0	Predicted: 1
Actual: 0	584 TN	449 FP
Actual: 1	43 FN	331 TP

Comment
Will “catch” more potentially churned customers, but will make more mistakes in predicting one as churned, increasing costs.



Logistic Regression - Top 10 Important Features



LightGBM - Top 10 Important Features

- The majority of the important features are, indeed, what we also found during the analysis
- Tenure, charges, short-term contracts and fiber optic service seem to be the main indicators

Recommendation

What is the best model?

To evaluate which model is better, let’s calculate the expected profit from each of them:

- Assume that a customer who renews his contract brings the company **65€ per month**
- The promotional activity (to prevent them from churning) has a cost of **1€**
- If a customer responds positively, then the net profit amounts to **64€ per month**
- The cost of non-response of the customer to the promotional campaign is **1€**

Case A: Logistic Regression

P(TP) = 299/1047 = 0.286

P(FP) = 292/1047 = 0.279

P(TN) = 741/1047 = 0.708

P(FN) = 75/1047 = 0.072

Expected Profit

0.286 * 64 - 0.279 * 1 = **18.03€**

Log. Reg.	**Predicted: 0**	**Predicted: 1**
Actual: 0	741 **TN**	292 **FP**
Actual: 1	75 **FN**	299 **TP**

Comment

If a company contacts customers labeled as potential churned, then it can expect a profit of about **18.03€ on average per customer.**

Case B: LightGBM

P(TP) = 331/1047 = 0.316

P(FP) = 449/1047 = 0.429

P(TN) = 584/1047 = 0.558

P(FN) = 43/1047 = 0.041

LightGBM	Predicted: 0	Predicted: 1
Actual: 0	584 TN	449 FP
Actual: 1	43 FN	331 TP

Comment

If a company contacts customers labeled as potential churned, then it can expect a profit of about **19.80€ on average per customer.**

Expected Profit

0.316 * 64 - 0.429 * 1 = **19.80€**

Thank you!