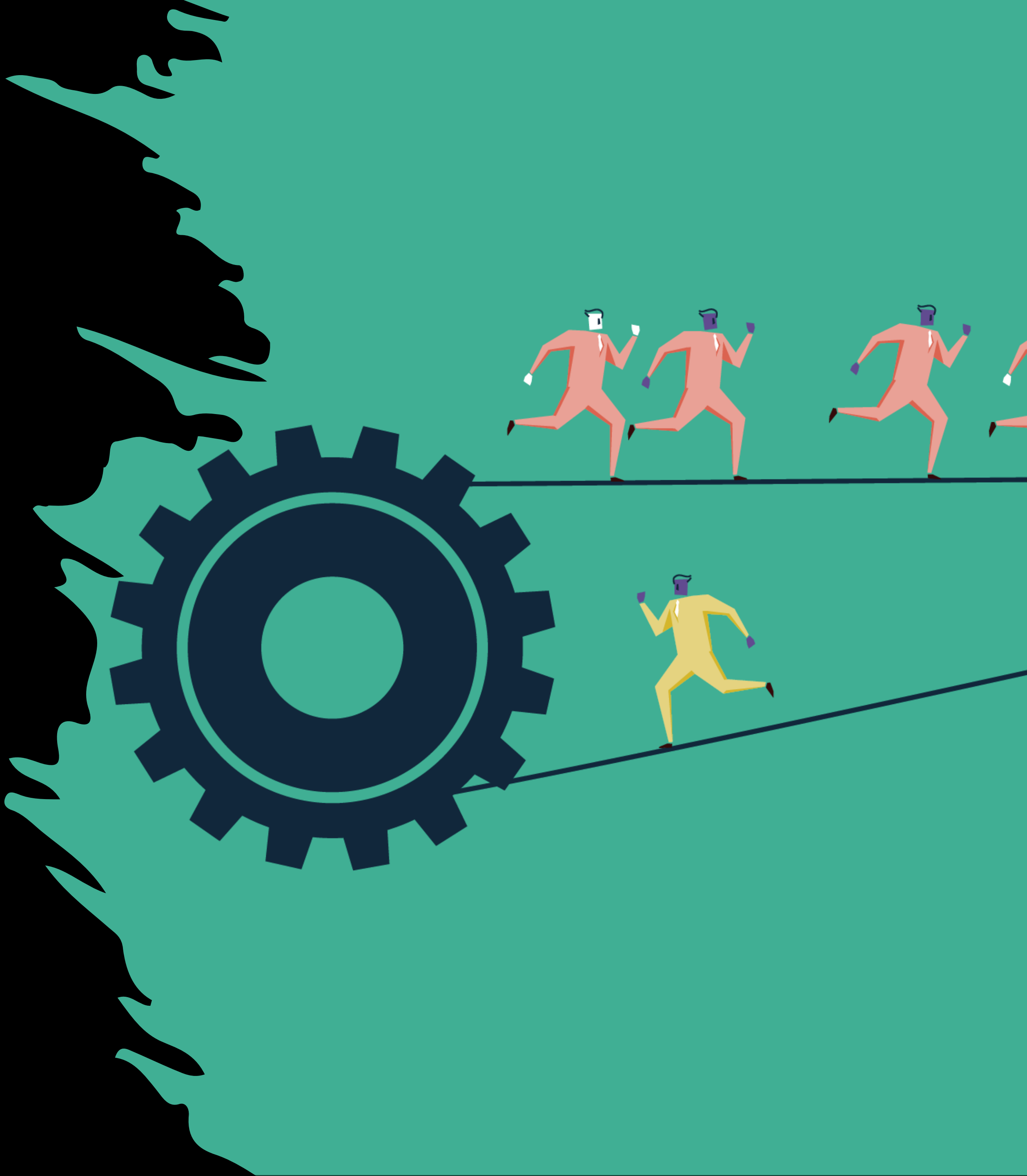
Telecom
Customer Churn

Improving customer retention with Machine learning



Agenda

01
Introduction
Explain the problem regarding Customer churn.

02
Data
Explain and describe the data.

03
Analysis
Exploratory Data Analysis.

04
Modeling
Describe modelling methodology and results.

05
Conclusion
Summary and recommendations.

01

INTRODUCTION

Explain the problem regarding Customer churn.

What is customer churn?

- Customer churn happens when customers decide to **stop using products or services** from an organization.
- It is a very important factor since it **costs 10 times** more to acquire new customers than it does to retain existing customers.
- Customer churn can prove to be a roadblock for an **exponentially growing organization**.
- Hence, a **retention strategy** should be decided.

Telecom vs Customer churn

- **Telecommunications companies** are usually **not the most popular companies** with consumers.

- People often express **frustration with the performance** of service providers.

- As a result, it is not surprising to learn that **telecommunications companies** have a **high customer churn rate**.

- Customer loyalty is the key to **profitability**.

- Therefore, finding factors that **increase customer churn** is important to take necessary actions to **reduce** this churn.

02

DATA

Explain and describe

the data.

Our Data

Numerical

Categorical

Quantitative Information

- Tenure
- Monthly Charges
- Total Charges

Customer Information

- Gender
- Partner
- Senior Citizen
- Dependents
- Etc.

Services Information

- Phone Service
- Internet Service
- Online Security
- Tech Support
- Etc.

Payment Information

- Contract
- Paperless Billing
- Payment Method

Descriptive Statistics

Total Customers

7032

Avg Monthly Charges

65€

Avg Tenure

32 months

Churn Rate

27%

No. of Contract Types

3 contracts

No. of Payment types

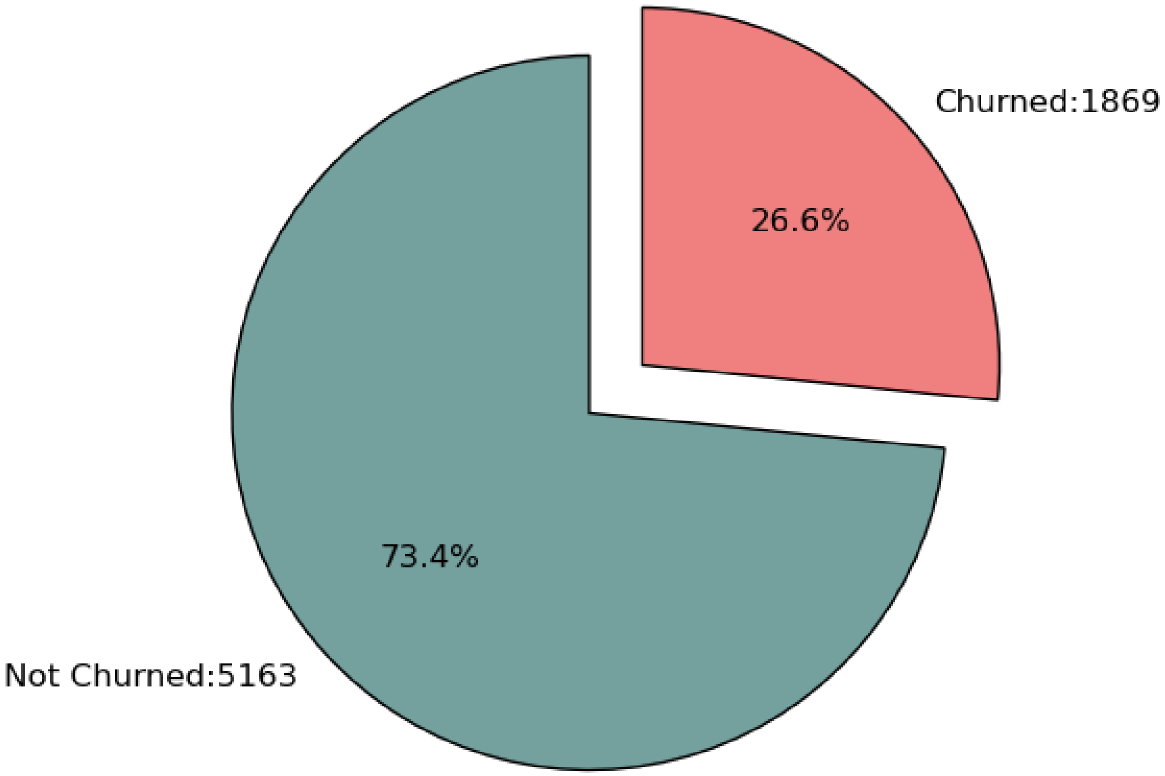
4 types

03

ANALYSIS

Exploratory Data Analysis.

Distribution of Churn



- As we can see from the pie chart, the dataset is **imbalanced** in a near about **3 to 1 ratio** for Not-Churn vs Churn customers.
- Due to this, **predictions** will be **biased towards Not-Churn** customers.
- Hence, we will try to **fix** later this issue using an **oversampling** technique named **SMOTE**.

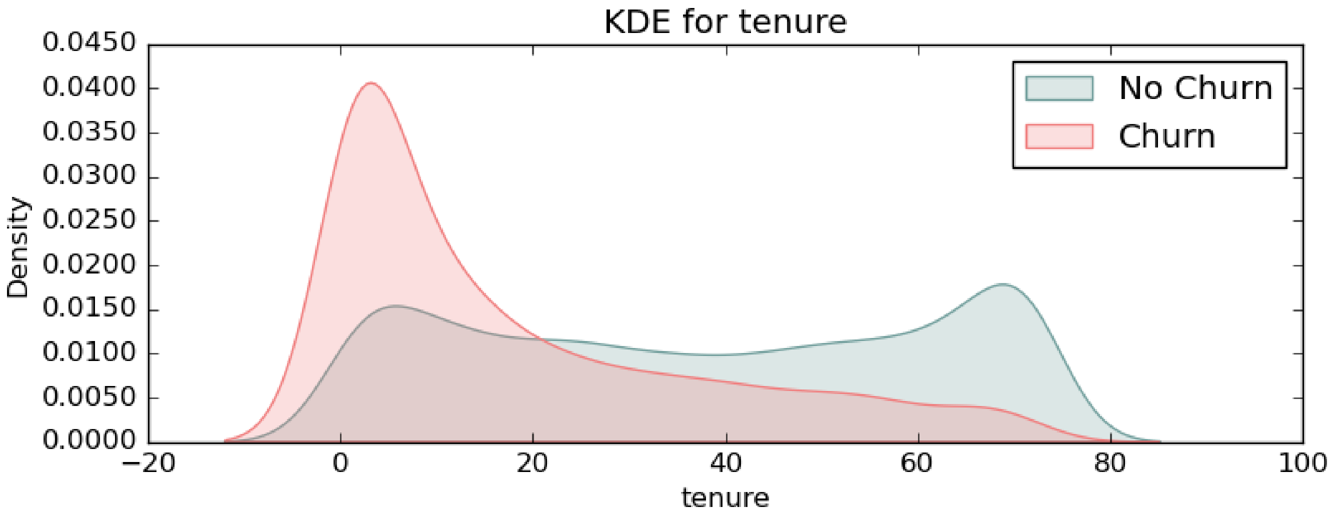
Numerical Features vs Churn

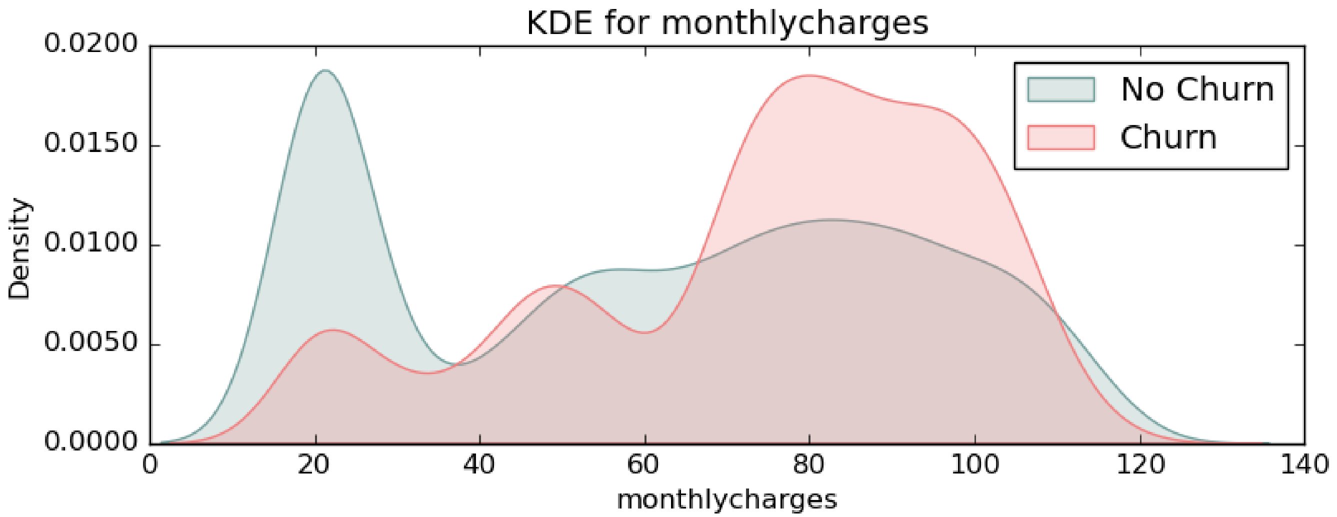
- Tenure and Monthly Charges kind of create a **bimodal distribution** with **peaks** present at **0 - 70** and **20 - 80**, respectively.

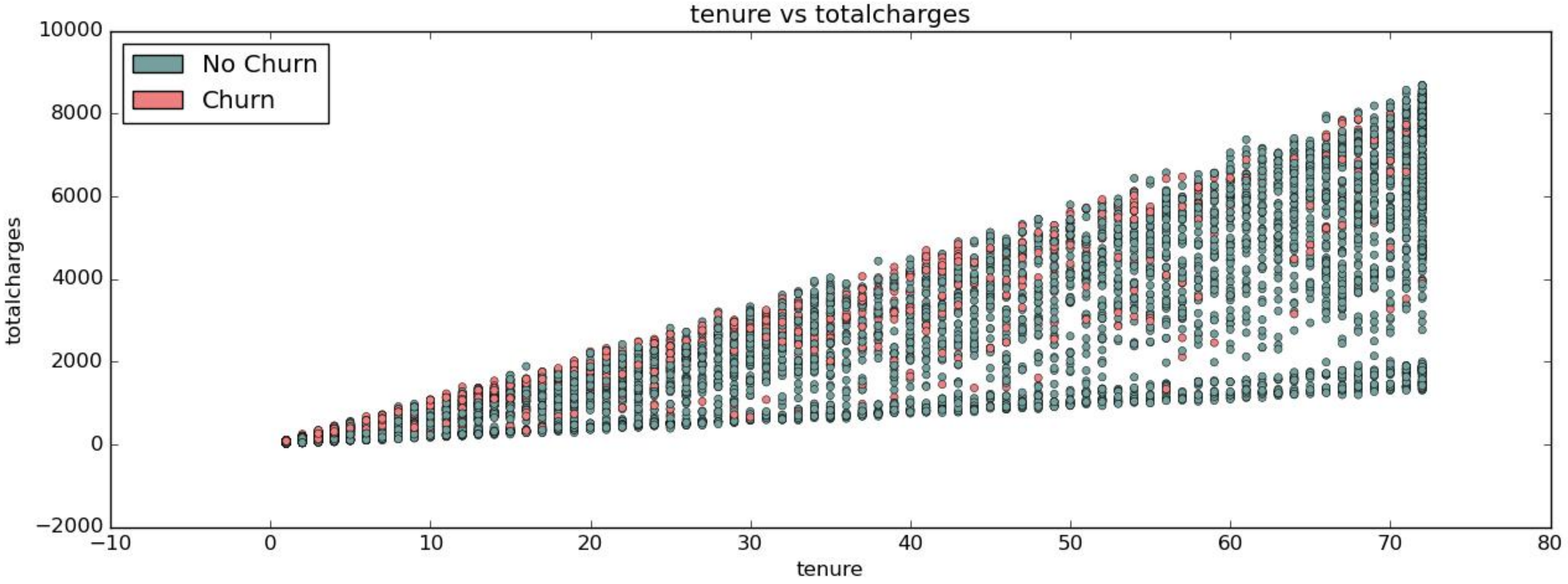
- Customers tend to **churn more** when :

  1. **tenure is between 1 - 5 months.**

  2. **Monthly charges range between 70-100 euros.**







- It is obvious that **as Tenure increases, Total Charges increases** as well.
- So which customers are **leaving**?
  1. The ones who are **charged the highest** of their **tenure period**.
  2. A few whose **Total Charge** rank in the **middle**.

Categorical Features vs Churn

- Since we have a lot of Categorical Variables in our dataset, we will **focus on** those that have the **highest or lowest churn rate**.

- For that purpose, we will set the **thresholds** below:
  1. **Higher than 30%** ~> tend to churn
  2. **Lower than 10%** ~> are loyal

Customer Information		**Not Churn**	**Churn**	**Churn Rate %**
Senior Citizen	Yes	666	476	**42%**
	No	4497	1393	24%
Partner	Yes	2724	669	20%
	No	2439	1200	**33%**
Dependents	Yes	1773	326	16%
	No	3390	1543	**31%**

- Nearly **1 out of 2 Senior Citizens** tend to Churn.

- Customers **with Partner** and those **with Dependents** are **less likely** to churn.

Categorical Features vs Churn

Services Information		Not Churn	Churn	Churn Rate %
Internet Service	Yes	3756	1756	**32%**
	No	1407	113	**7%**
Fiber Optic	Yes	1799	1297	**42%**
	No	3364	572	15%
Online Security	Yes	1720	295	15%
	No	3443	1574	**31%**
Tech Support	Yes	1730	310	15%
	No	3433	1559	**31%**

- Customers with **no Internet Service** does **not** seem to **churn**.
- Customers who use **Fiber Optic** as Internet Service **tend to Churn**.
- A **high number of customers** have **switched their service provider** when it comes down **poor services**.
- Especially, regarding **Online Security** and **Tech Support**, **1 out of 3** customers tend to **drop out**.

Categorical Features vs Churn

Payment Information		Not Churn	Churn	Churn Rate %
Contract	Month-to-Month	2220	1655	**43%**
	One year	1306	166	11%
	Two year	1637	48	**3%**
Paperless Billing	Yes	2768	1400	**34%**
	No	2395	469	16%
Electronic Check	Yes	1294	1071	**45%**
	No	3869	798	17%

- It is obvious that customers with **short-term** contract seems to **churn more**.
- Moreover, nearly **1 out of 2** customers having **month-to-month contract** are **leaving** the company.
- **Paperless Billing** displays a high number of customers being **churned** out, **more than 1 out of 3** customers.
- Although the customers **pay more** with **Electronic check**, the **churn rate** is **higher** compared to the other types.

Summary

- The dataset is **not balanced**. Therefore, we need to take care of this issue after **training** the model.
- Customers tend to **churn** when the **tenure** is between **1-5 months** and **Monthly Charges** range between **70-100 euros**.
- In addition, customers with **short-term** contract seems to **churn more.**
- **1 out of 2 Senior Citizens** seems to leave the company compared to those with a **Partner or Dependents,** who seem to be more loyal.
- Consumers who use **Fiber Optic** tend to **opt out** the company more.
- Moreover, **Paperless Billing and Electronic Payment** display a **high number** of customers being **churned** out.

04

MODELING

Describe modelling methodology and results.

Modeling Process

01

Data Pre-processing

- Data Exploration
- Data Cleansing
- One-hot Encoding

02

Feature Engineering

- Feature Selection using ANOVA & Chi-squared test
- Feature Scaling

03

Class Imbalance

- Oversampling minority class using SMOTE

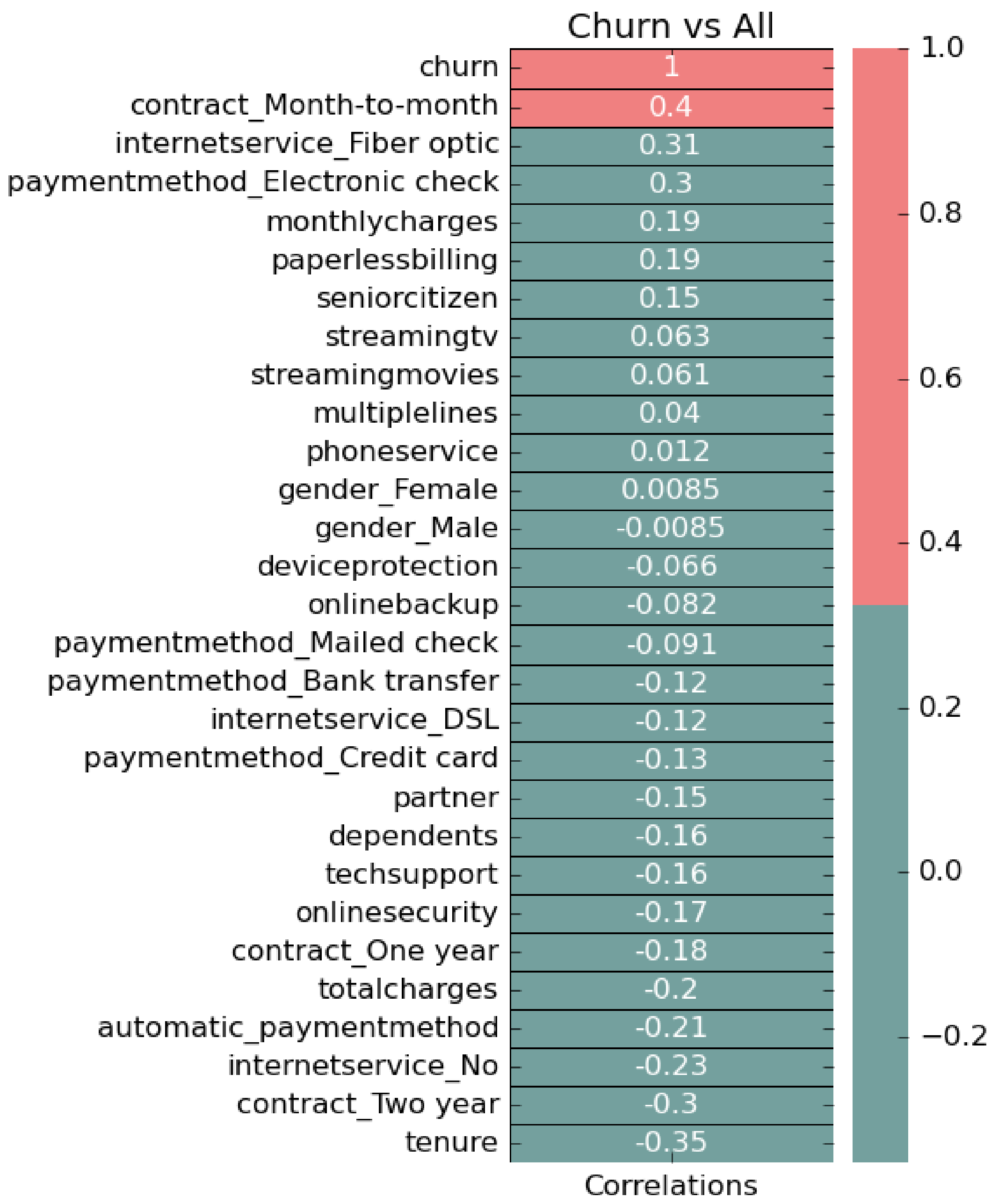
04

Models & Evaluation

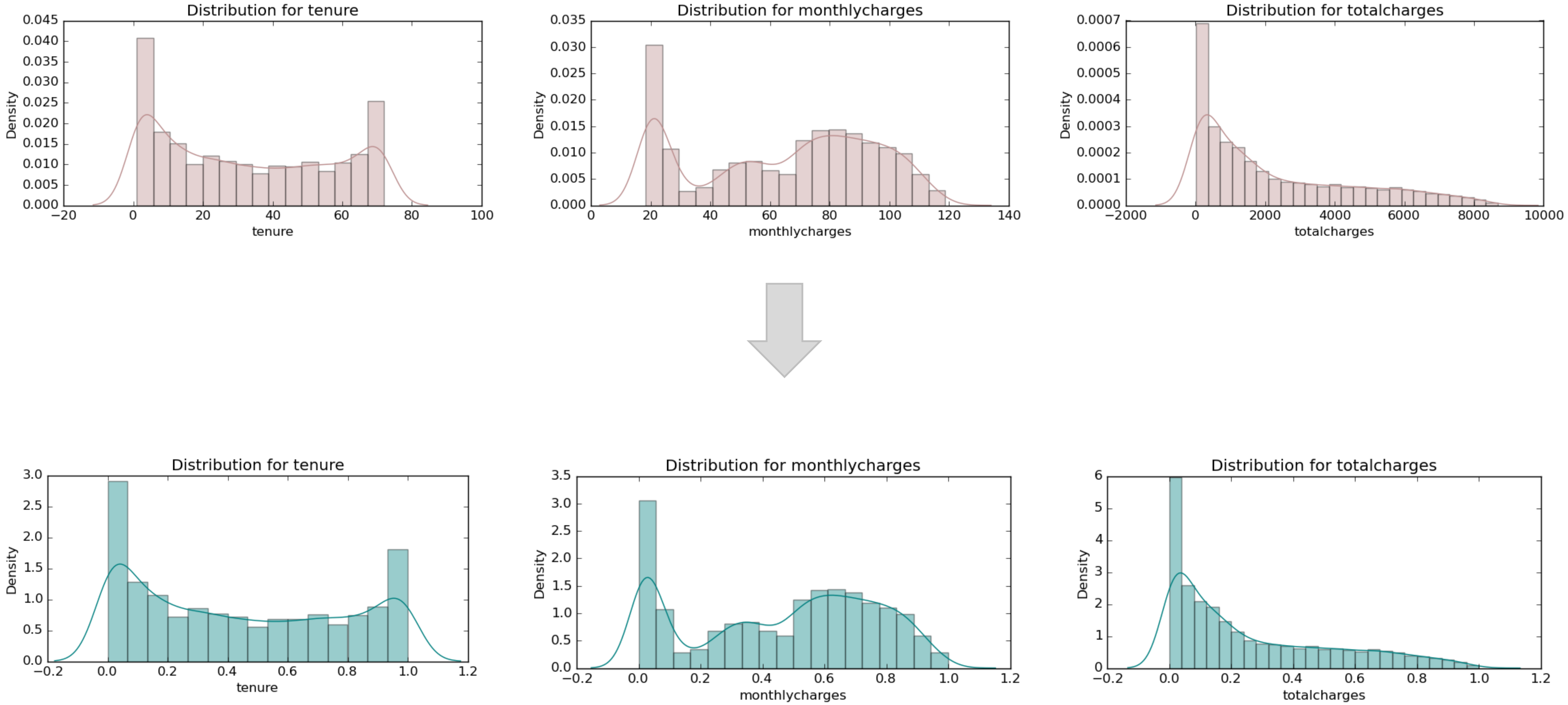
- Cross Validation
- Hyperparameter Tuning
- Modeling
- Feature Importance

Feature Selection

- We **dropped** the features with correlation **coefficient between (-0.1 , 0.1)**.
- Moreover, **Chi-squared Test** for Categorical Variables and **ANOVA test** for Numerical Variables showed the same result.

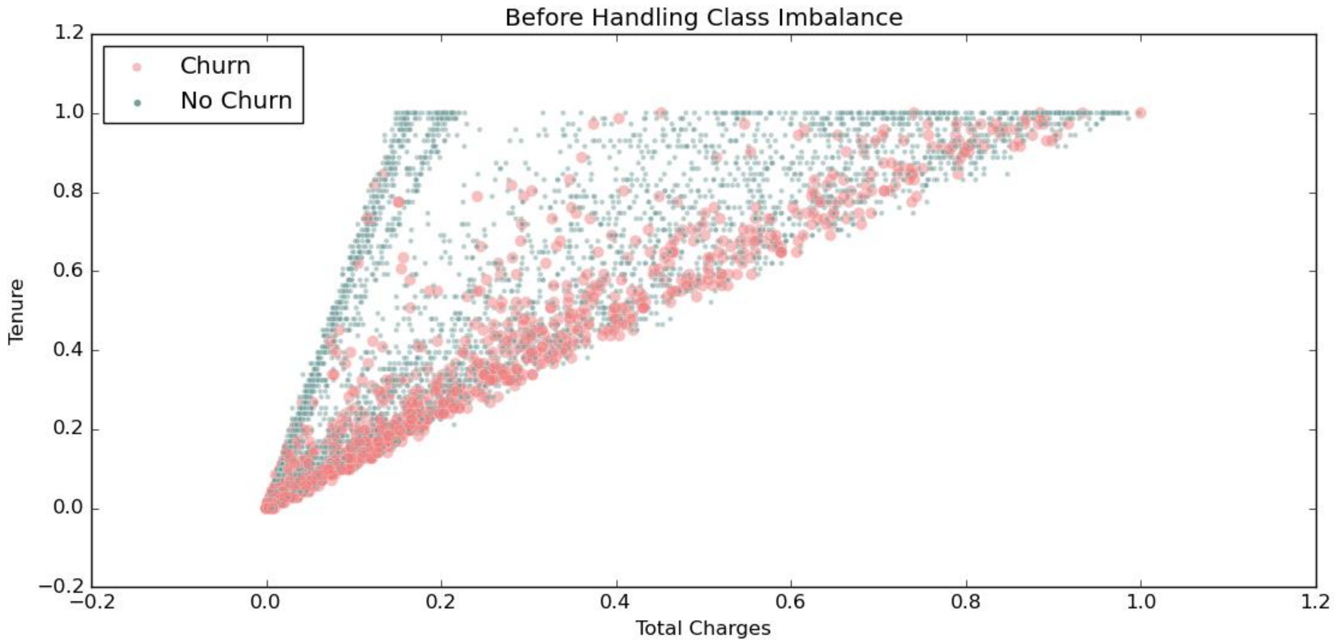


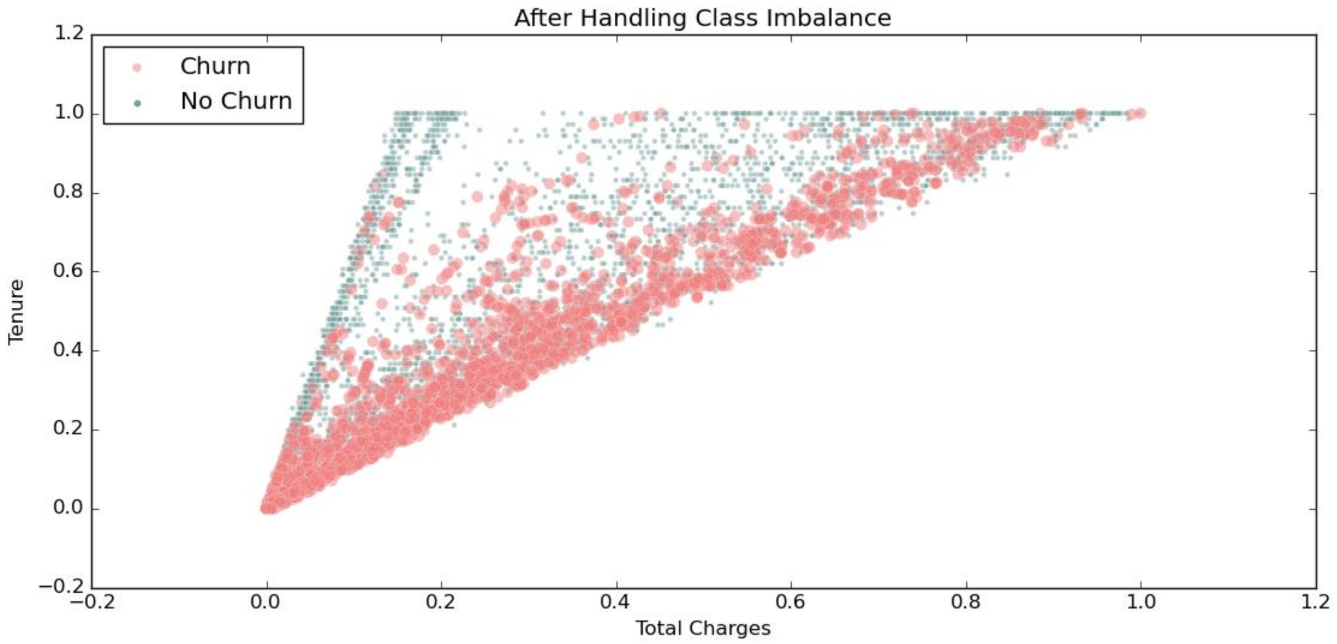
Feature Scaling (Numerical Variables)



Class Imbalance

- As we mentioned before, The dataset is **imbalanced**.
- Therefore, we fixed this issue after **training** the model to avoid changing the test set.
- We used an **oversample** technique called **SMOTE**.
- SMOTE **oversample** the data in the **minority class** by finding its **k-nearest neighbors**.





Models & Evaluation

Models	Precision	Recall	F1 Score	ROC	Accuracy
Logistic Regression	0.52	0.78	**0.62**	0.76	0.75
Kernel SVM	0.51	0.77	0.61	0.75	0.74
Gaussian NB	0.49	0.80	0.61	0.75	0.73
Adaboost	0.50	0.79	0.61	0.75	0.73
Gradient boost classifier	0.50	0.74	0.6	0.74	0.74
SVC	0.46	**0.84**	0.59	0.74	0.69
XGBoost Classifier	**0.54**	0.66	0.59	0.73	0.76
Random Forest	0.53	0.59	0.56	0.70	0.75
KNN	0.48	0.64	0.55	0.69	0.72
Decision Tree Classifier	0.46	0.52	0.49	0.65	0.71

** Results are sorted by F1 Score, Recall, Accuracy*

Models & Evaluation

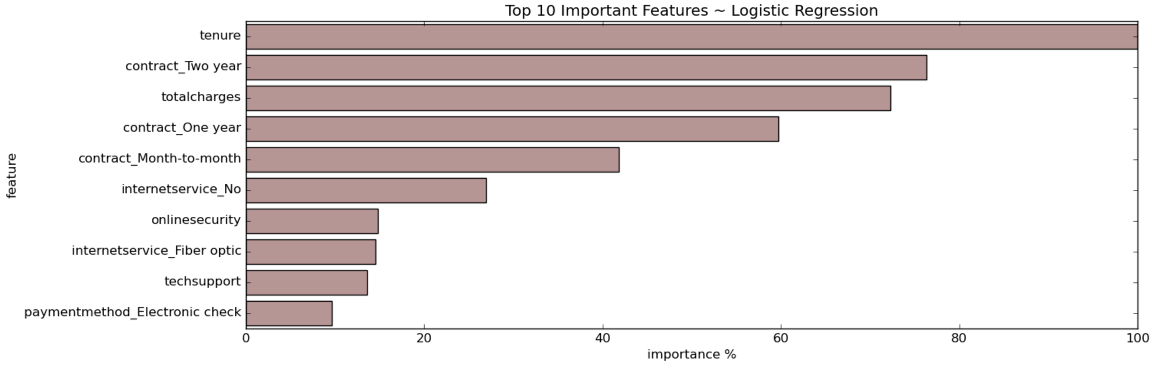
Let’s summarize the **3 best models** based on **F1 Score, Recall & Precision**:

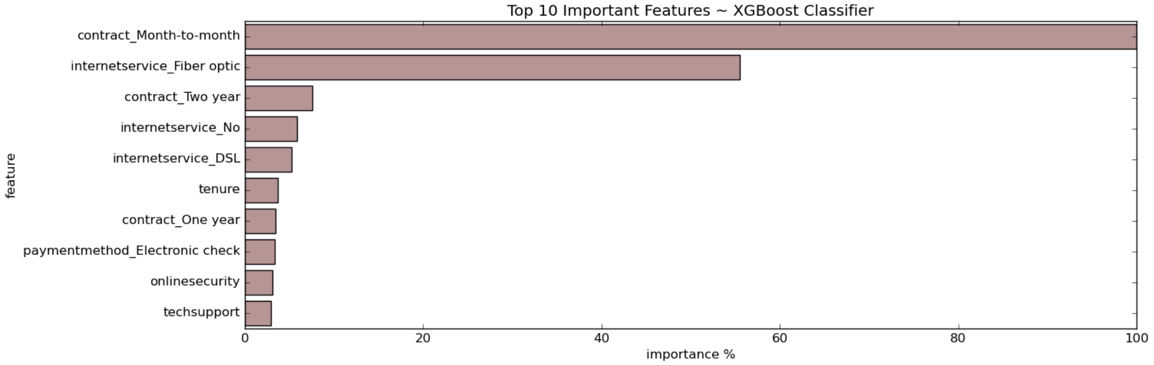
- **Logistic Regression** has the **best F1-score** compared to the other models we used. This means that it has the **best overall performance** in order to **balance precision** and **recall**.

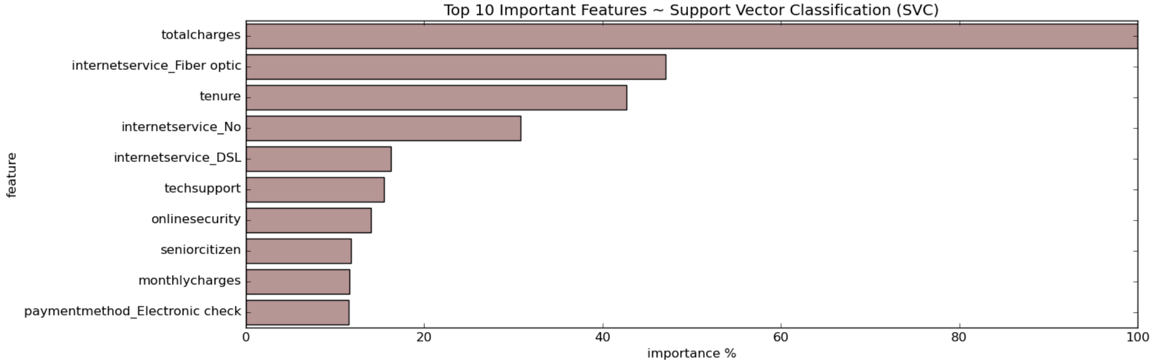
- **Support Vector Classifier** has the **best Recall score** which means that the model **predict more accurate the customers that had churned**.

- **XGBoost Classifier** has the **best Precision score** which means that the model **determines** more efficient **how many** of the predicted customers **drop out**. On the other hand, this model is more willing to accept a prediction with less proof which will **lead us to predict some customers as churned but they are loyal**.

Future Importance







The **important features** are almost the same for all 3 models and confirm our findings in EDA where **Tenure**, **short-term Contracts**, **Fiber Optic** and **Electronic check** are the most **important**.

05

CONCLUSION

Summary and
recommendations.

What model the company should choose?

- In order to answer efficiently this question, we are going to **simulate** a **Financial Cost-Effective Strategy** that applies 2 times higher costs to False Negative than to False Positives.

Let's make some assumptions here:

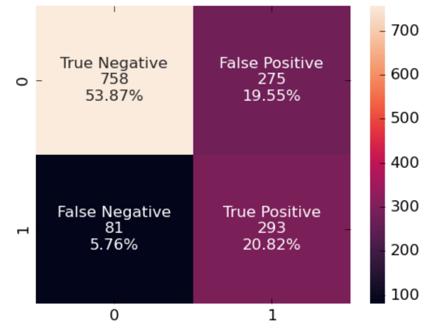
- The **average customer acquisition cost** for telecom is **300** euros.
- Assign to the **true negatives** the cost of **0 euros, since** the model correctly identified a happy customer.
- **False negatives** are the **most problematic**, because they incorrectly predict that a churning customer will stay so we will assign the value of **300 euros**.
- Finally, for customers that the **model identifies as churning**, we will assume a retention incentive in the amount of **150 euros.** This is the cost of **both true positive and false positive** outcomes. In the case of false positives (the customer is happy, but the model mistakenly predicted churn), we will assign the 150 euros concession.

To understand which is the best model for the company to use to reduce its costs we should **minimize a cost function** that looks like this:

Cost = 300FN(C) + 0TN(C) + 150FP(C) + 150TP(C) , where C:Count

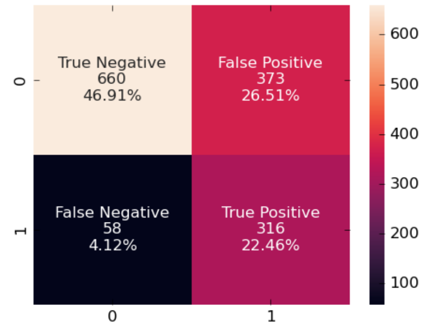
Logistic Regression:

Cost_LR = 300 * 81 + 0 * 758 + 150 * 275 + 150 * 293 = **109,500 euros**



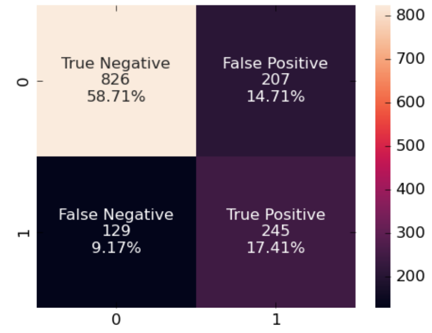
Support Vector Classifier:

Cost_SVC = 300 * 58 + 0 * 660 + 150 * 373 + 150 * 316 = **120,750 euros**



XGBoost Classifier:

Cost_XGBC = 300 * 129 + 0 * 826 + 150 * 207 + 150 * 245 = **106,500 euros**



THANK YOU!!