



# Final Project

Telco Customer Churn Prediction

---

**Adhang Muntaha Muhammad**





01

# **BUSINESS & DATASET UNDERSTANDING**

# Dataset Information



- Customer information from a **telco company**
- This company provides various services such as **streaming, internet,** and **phone** service

# Attribute Information

Identifier	Target variable	Demographic information	Account information	Subscribed service
Customer ID	Churn status	Gender	Tenure	Phone service
		Senior citizen	Monthly charges	Multiple lines
		Partner	Total charges	Internet service
		Dependents	Contract	Online security
			Paperless billing	Online backup
		Payment method	Device protection	
			Tech support	
			Streaming TV	
		Streaming movies		

# Company Goals



- **Acquiring new customers** as much as we can
- **Retaining existing customers** as much as we can

# Cost Problems



Acquiring a new customer can be **25x more expensive**

Reference:

<https://hbr.org/2014/10/the-value-of-keeping-the-right-customers>

<https://www.outboundengine.com/blog/customer-retention-marketing-vs-customer-acquisition-marketing>

# Objectives



- **Predict** whether the customer will still use our service or **will leave** our service
- Understanding **customer behavior**
  - What keeps customers using our service
  - What makes customers leave our service



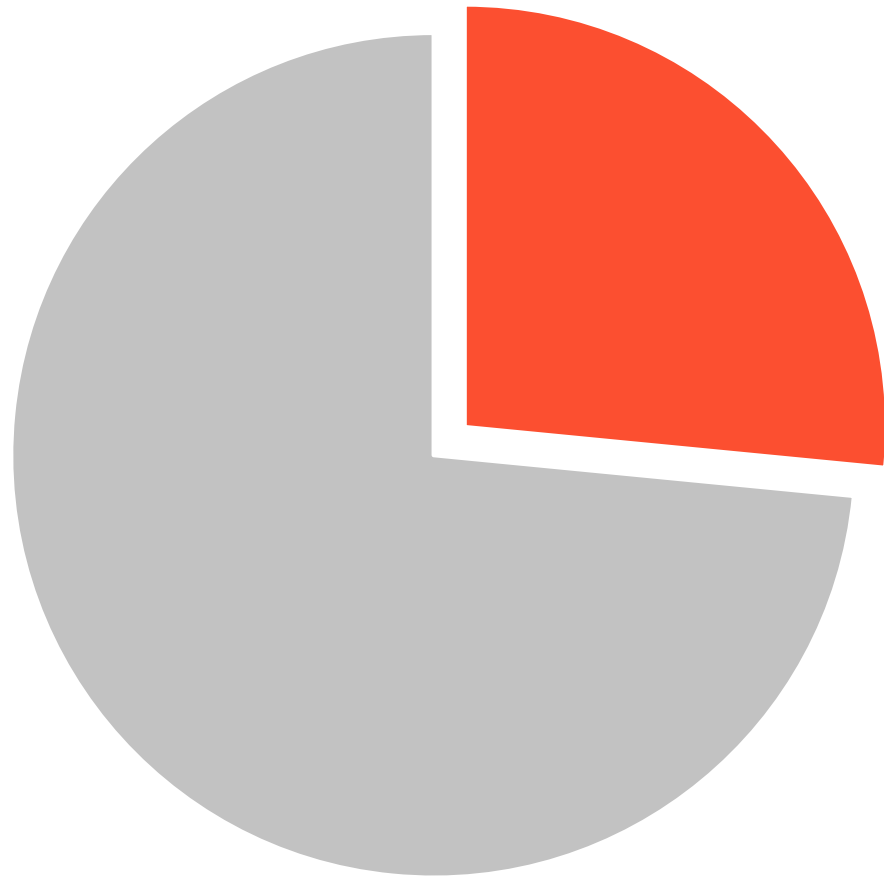
02



**EXPLORATORY  
DATA ANALYSIS**



# What Happened?



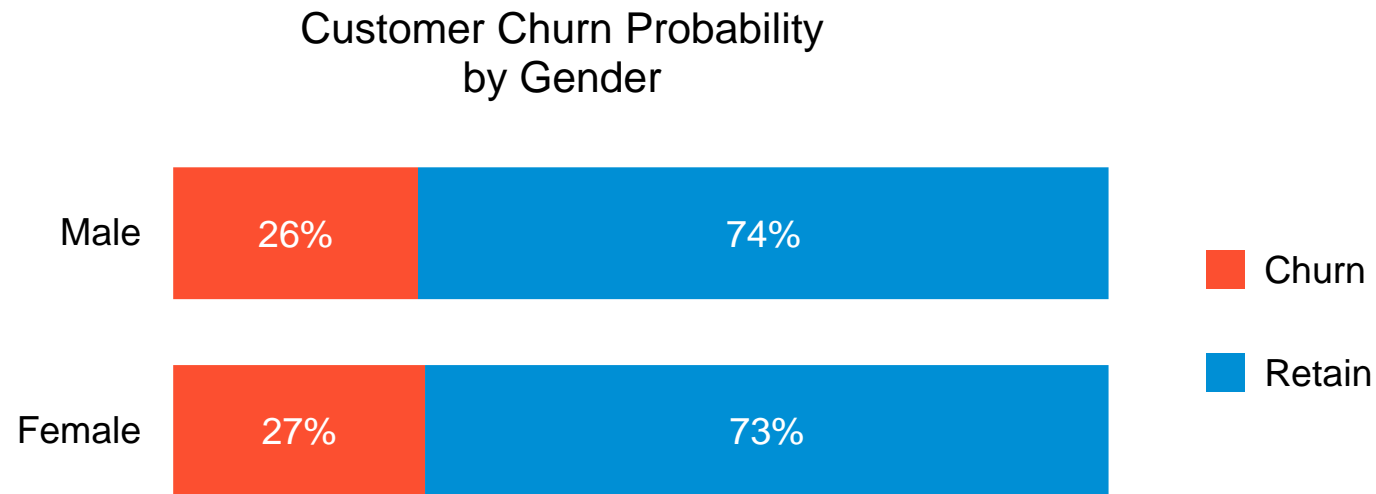
**27%**

Customers leave us! 😞

Technically speaking,  
This is an imbalanced dataset

# Why Did It Happen?

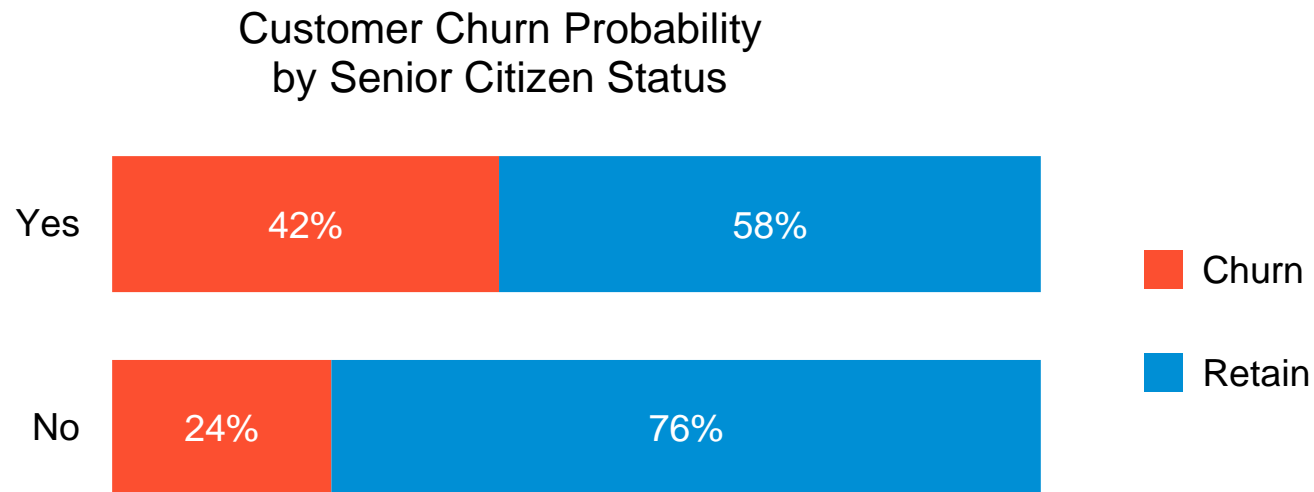
Not all attributes have a **strong relationship** with churn status



Both males and females almost have the same probability to churn  
We can say that customer's gender **has no relationship** with their churn status

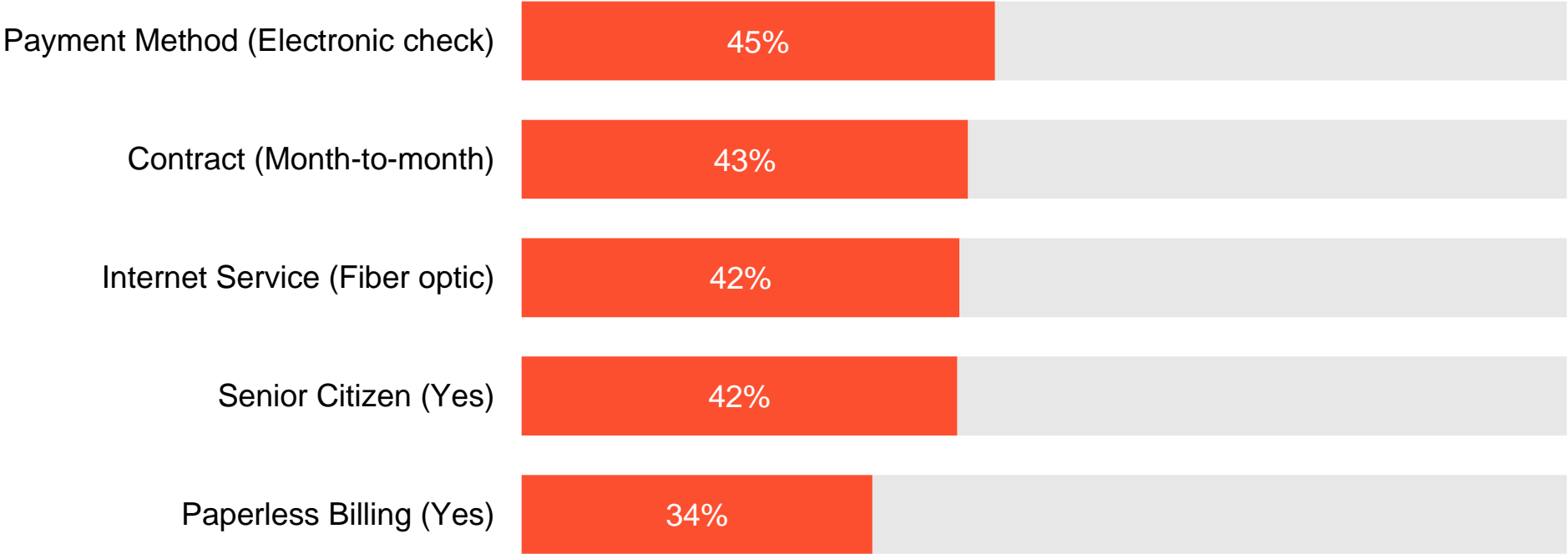
# Why Did It Happen?

Not all attributes have a **strong relationship** with churn status

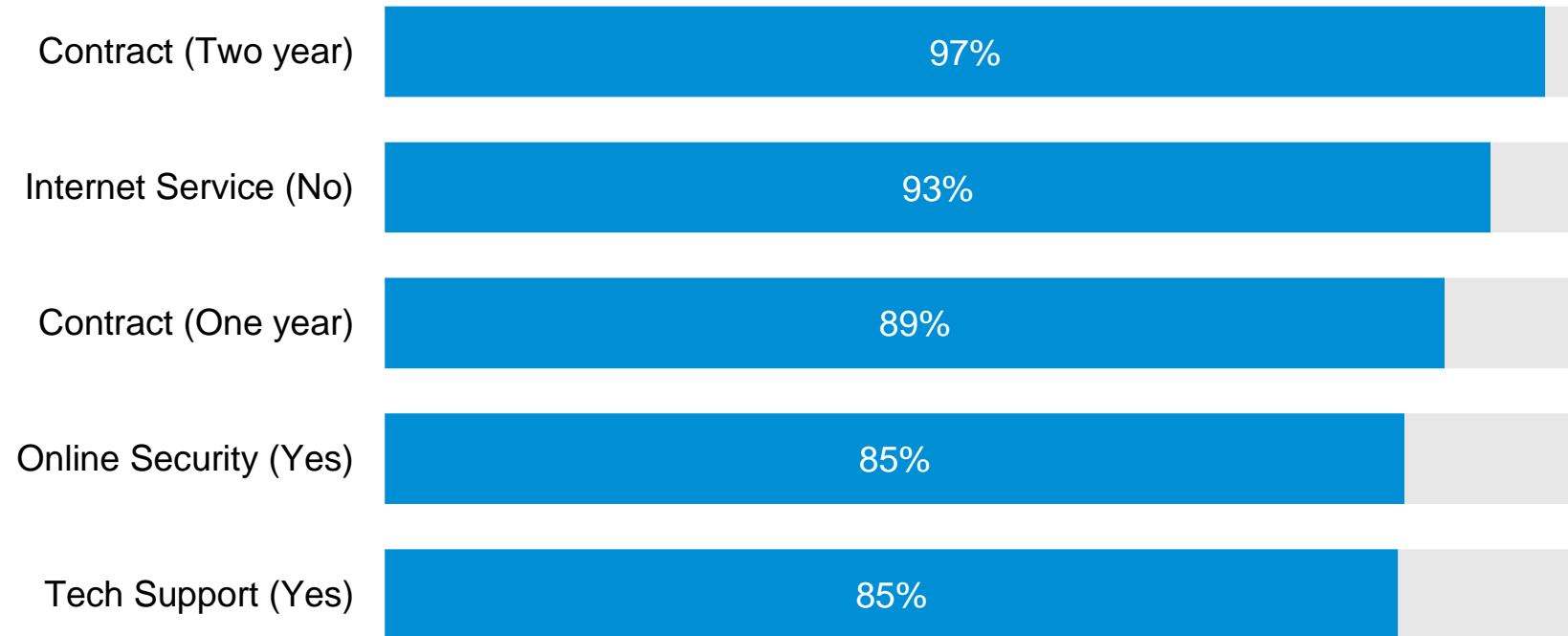


Senior citizens have a higher probability of churn than younger citizens  
We can say this attribute **has a relationship** with churn status

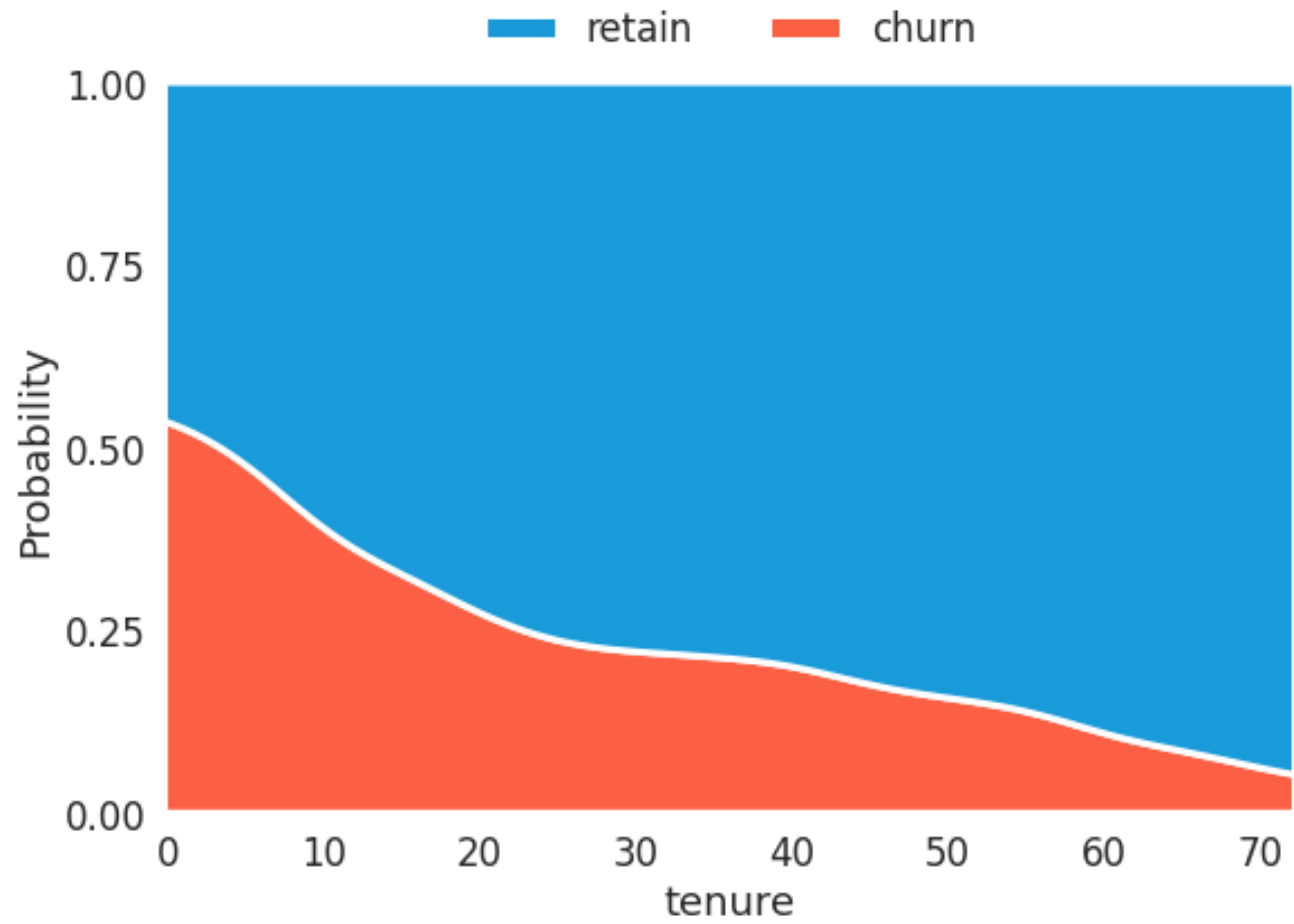
# Top 5 Churn Probability



# Top 5 Retain Probability



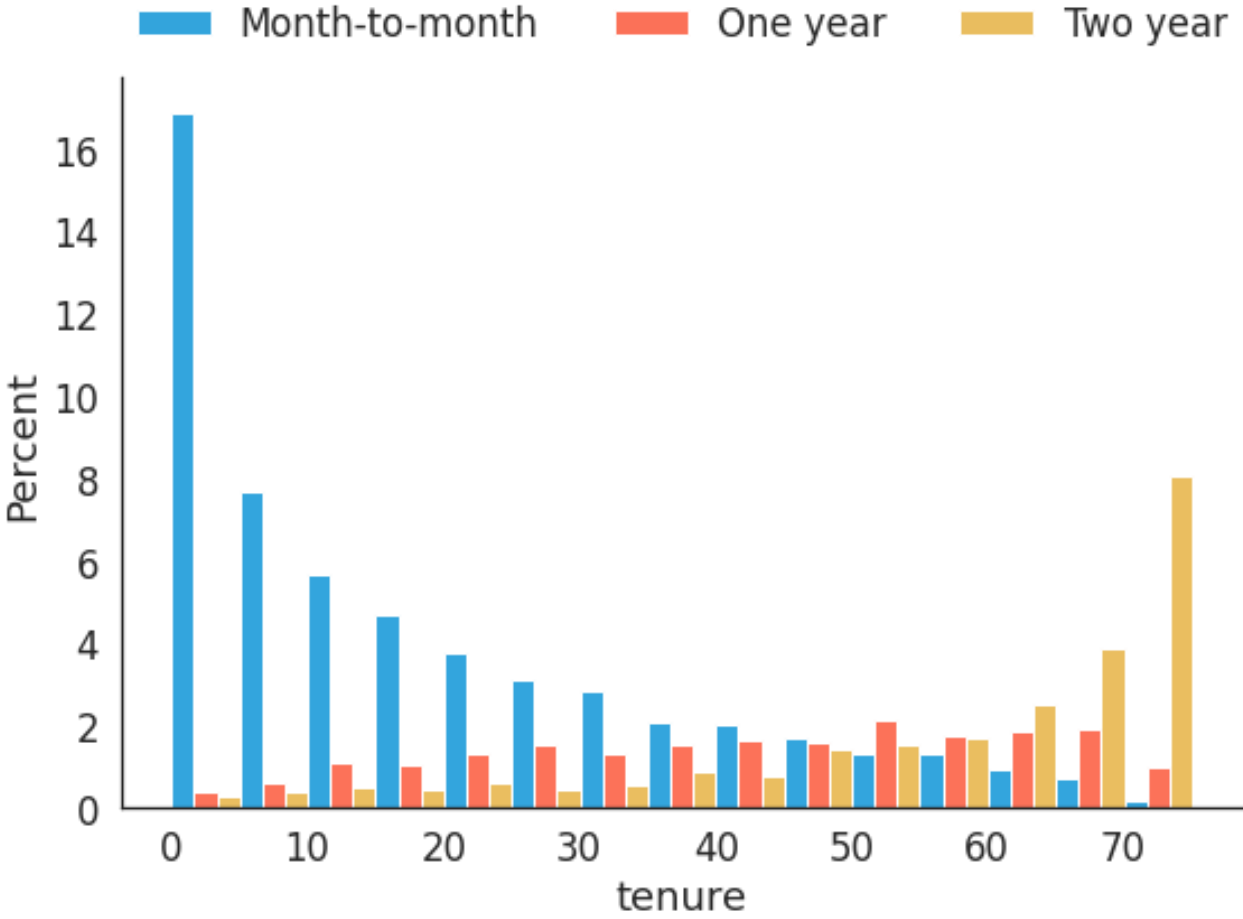
# Tenure



It has a clear trend!

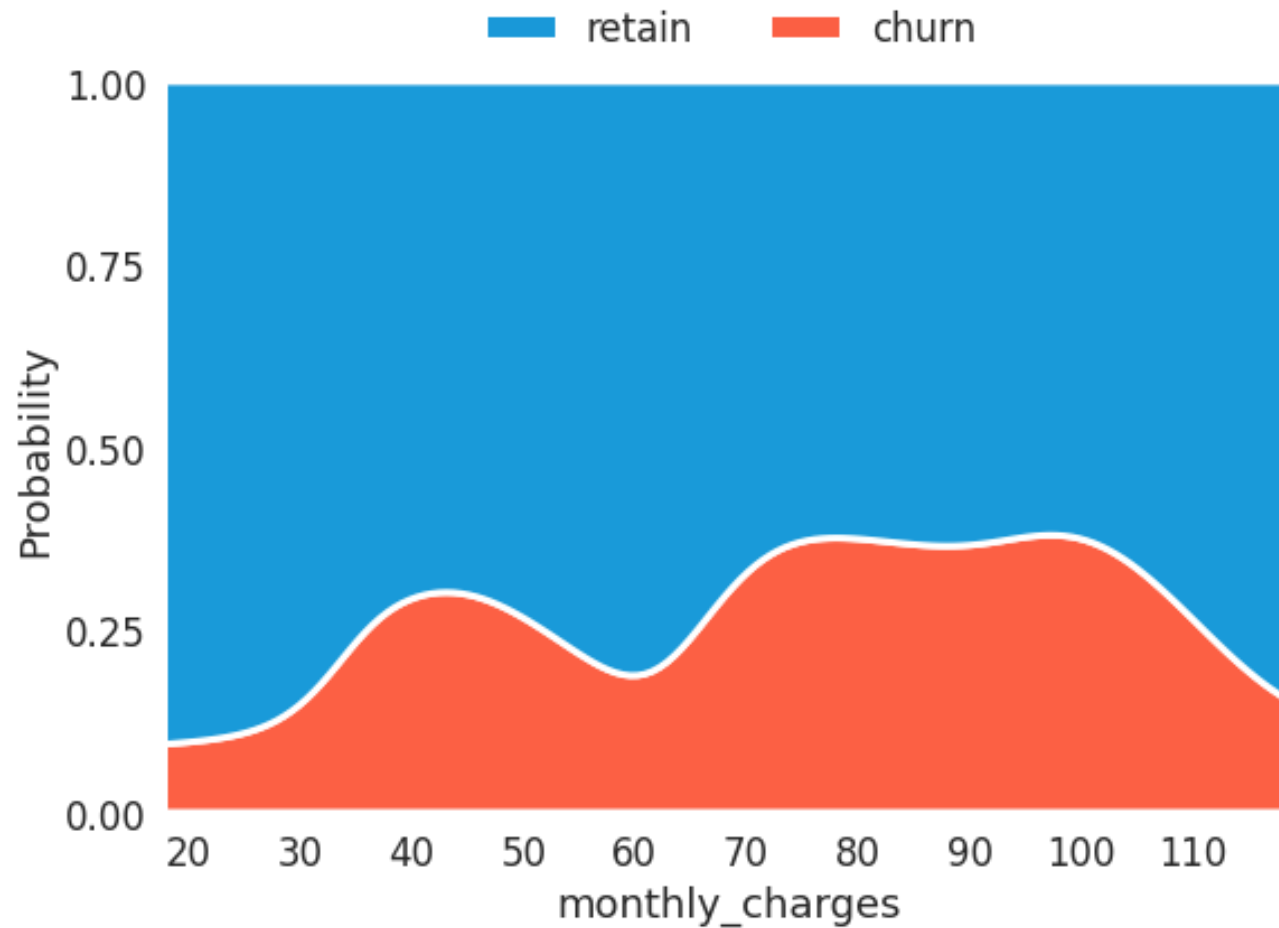
**Negative correlation**  
with the probability of  
churn

# Tenure by Contract



Customers with **short tenure** are more likely to have a **month-to-month** contracts

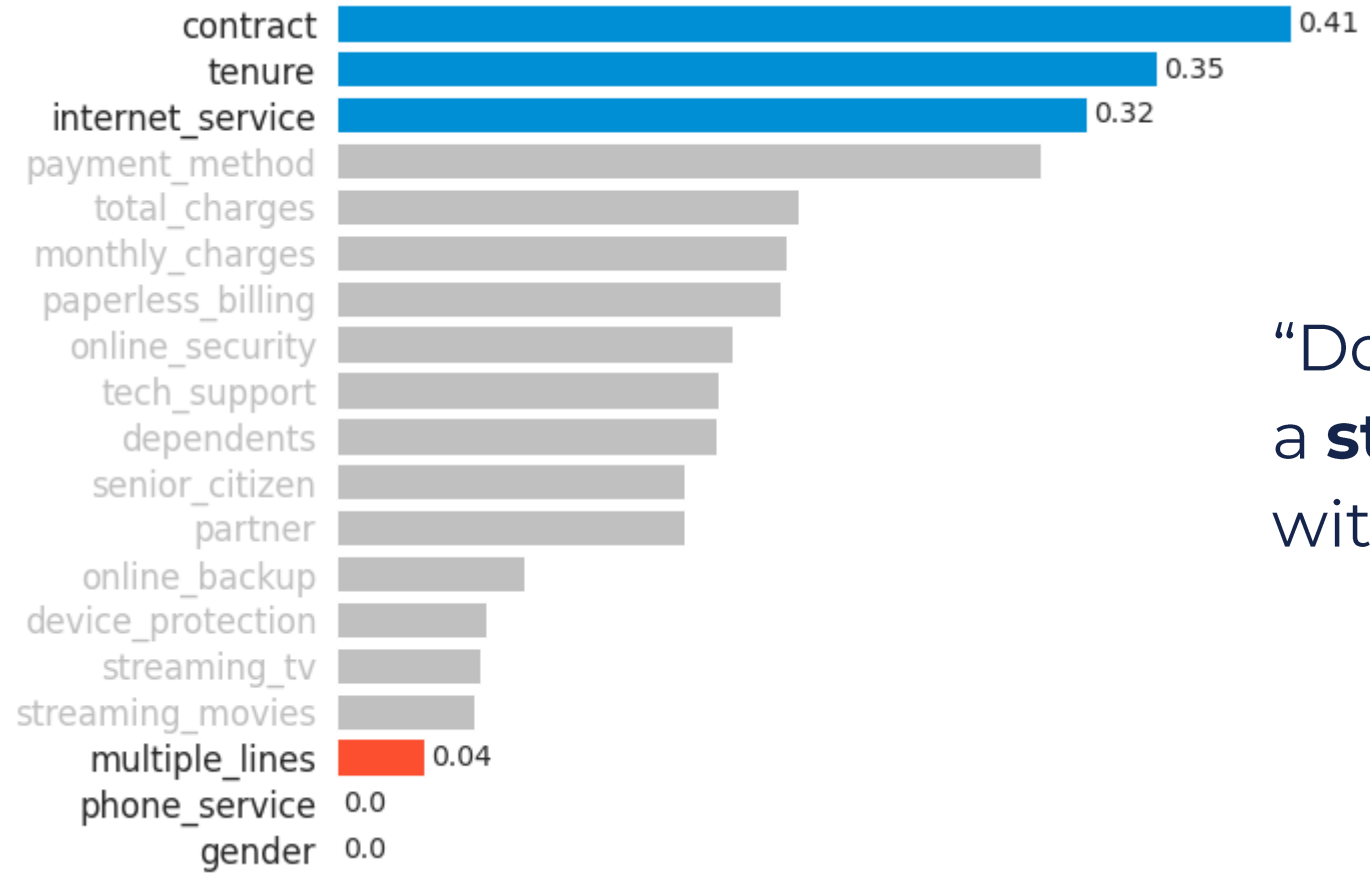
# Monthly Charges



It has no clear trend



# Attribute Associations to Churn Status



“Does this attribute have a **strong relationship** with churn status?”

04

**DATA  
PREPROCESSING**

# Missing Values

	tenure	total_charges	churn
488	0		No
753	0		No
936	0		No
1082	0		No
1340	0		No
3331	0		No
3826	0		No
4380	0		No
5218	0		No
6670	0		No
6754	0		No



	tenure	total_charges	churn
488	0	0	No
753	0	0	No
936	0	0	No
1082	0	0	No
1340	0	0	No
3331	0	0	No
3826	0	0	No
4380	0	0	No
5218	0	0	No
6670	0	0	No
6754	0	0	No

# Redundant Values

Attribute	Data variation
multiple_lines	No, Yes, <b>No phone service</b>
online_security	No, Yes, <b>No internet service</b>
online_backup	No, Yes, <b>No internet service</b>
device_protection	No, Yes, <b>No internet service</b>
tech_support	No, Yes, <b>No internet service</b>
streaming_tv	No, Yes, <b>No internet service</b>
streaming_movies	No, Yes, <b>No internet service</b>



Attribute	Data variation
multiple_lines	No, Yes
online_security	No, Yes
online_backup	No, Yes
device_protection	No, Yes
tech_support	No, Yes
streaming_tv	No, Yes
streaming_movies	No, Yes

## Replace to “No”

Avoiding multicollinearities & reducing the dimension

# Train – Test Split

Train : Test  
**70% : 30%**

	Variable	Shape
Original	X	(7043, 19)
	Y	(7043, )
Train set	X_train	(4930, 19)
	y_train	(4930, )
Test set	X_test	(2113, 19)
	y_test	(2113, )

# Feature Encoding

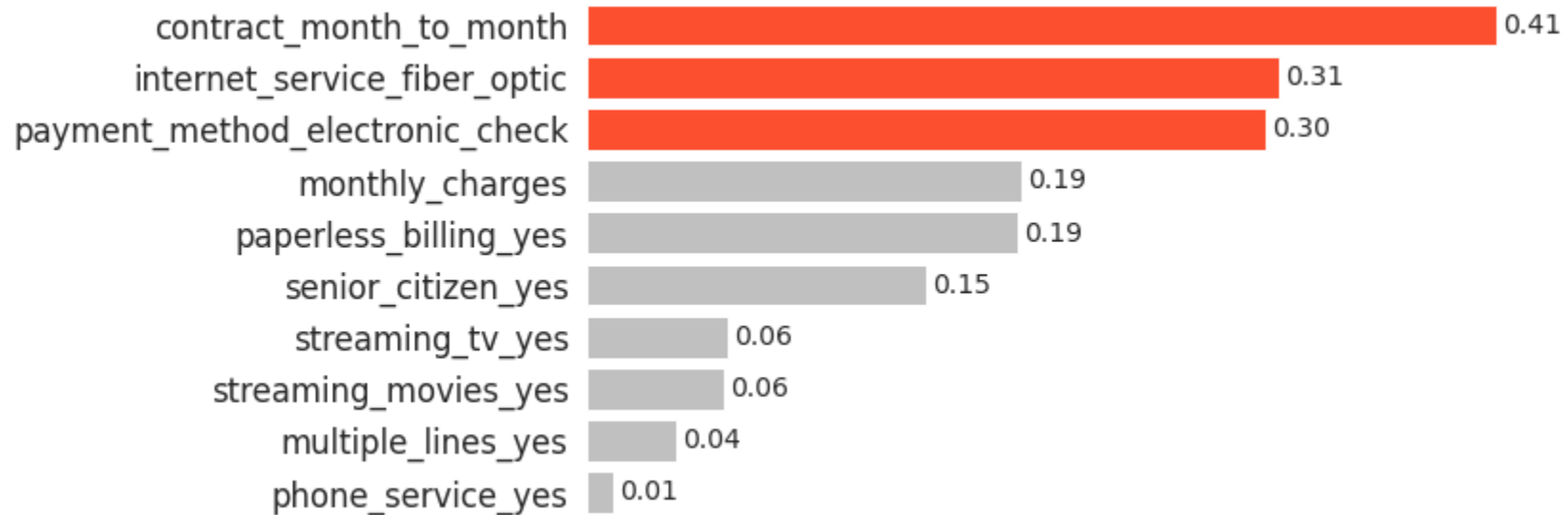
		6427	6971	96	5640
input	internet_service	No	Fiber optic	DSL	Fiber optic
output	churn	No	Yes	No	Yes



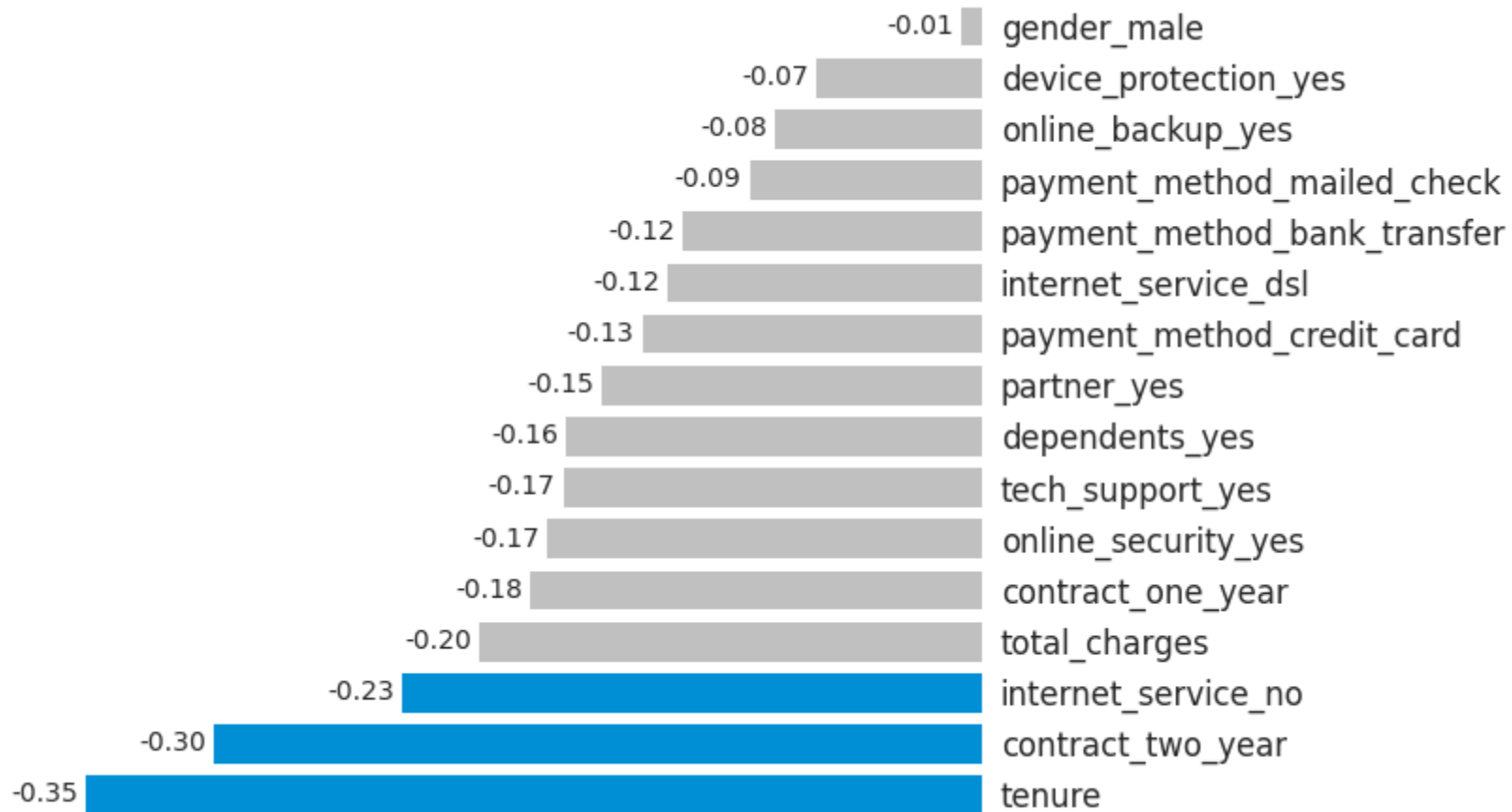
**LabelEncoder for target variable**  
**OneHotEncoder for input variables**

		6427	6971	96	5640
input	internet_service_dsl	0.0	0.0	1.0	0.0
input	internet_service_fiber_optic	0.0	1.0	0.0	1.0
input	internet_service_no	1.0	0.0	0.0	0.0
output	churn	0.0	1.0	0.0	1.0

# Positive Correlation to Churn Status



# Negative Correlation to Churn Status





# Feature Scaling

	6427	6971	96	5640
tenure	41.00	18.00	71.00	1.0
monthly_charges	20.15	99.75	66.85	79.6
total_charges	802.35	1836.25	4748.70	79.6



**MinMaxScaler**  
(range 0-1)

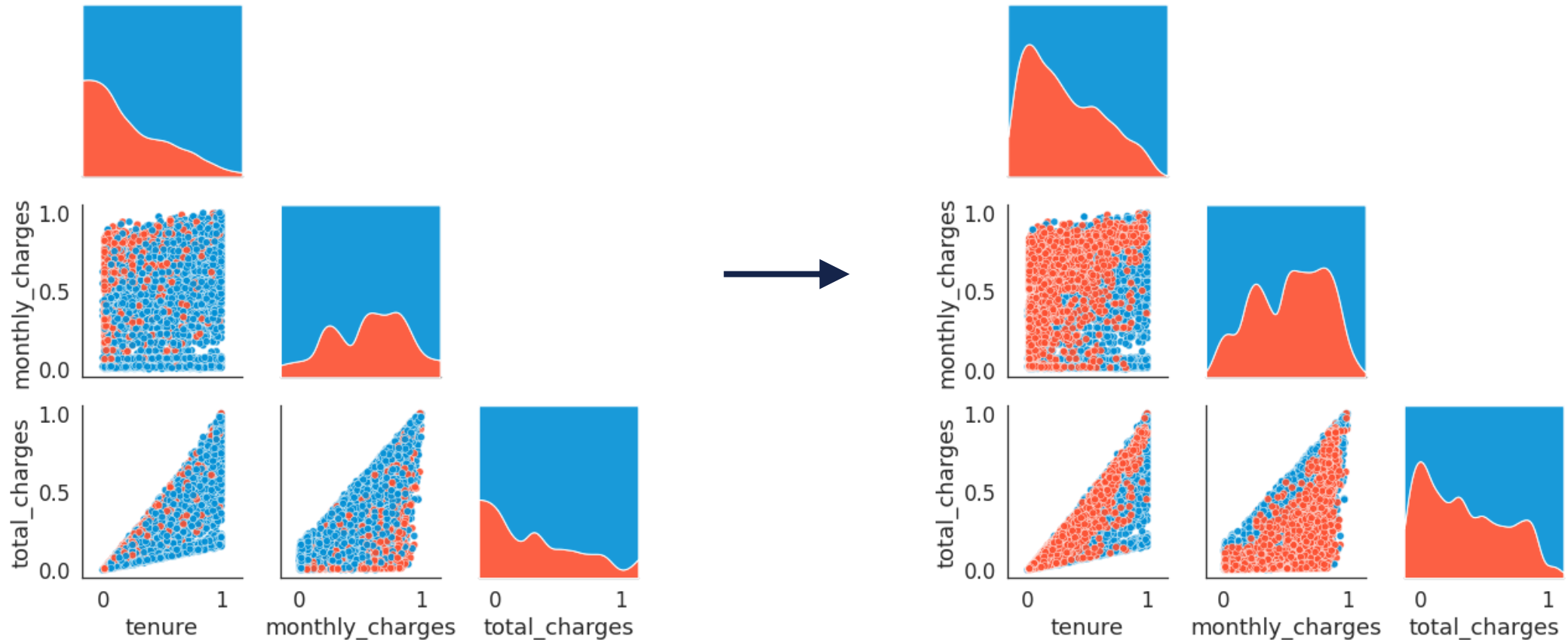
	6427	6971	96	5640
tenure	0.57	0.25	0.99	0.01
monthly_charges	0.02	0.81	0.48	0.61
total_charges	0.09	0.21	0.55	0.01

# Oversampling with SMOTE

Before SMOTE

After SMOTE

■ Churn ■ Retain





05



**MODEL  
DEVELOPMENT**

# Model List

Logistic Regression

Ridge Classifier

KNN

SVC

Neural Network

Decision Tree

Random Forest

Gradient Boosting Classifier

AdaBoost Classifier

CatBoost Classifier

Histogram Gradient Boosting


XGBoost

LightGBM

# SMOTE vs ADASYN

The average metrics of all models using default parameter

	SMOTE	ADASYN
Accuracy	0.760	0.752
Precision	0.707	0.704
Recall	0.732	0.731
F1 Score	0.713	0.707
ROC AUC	0.732	0.731



Note:  
Precision, recall, F1 score, and ROC AUC score are calculated using **macro average**

SMOTE has a higher performance than ADASYN,  
So, for the next step, I will **only use SMOTE** for simplicity reason

# Metrics using Default Parameter

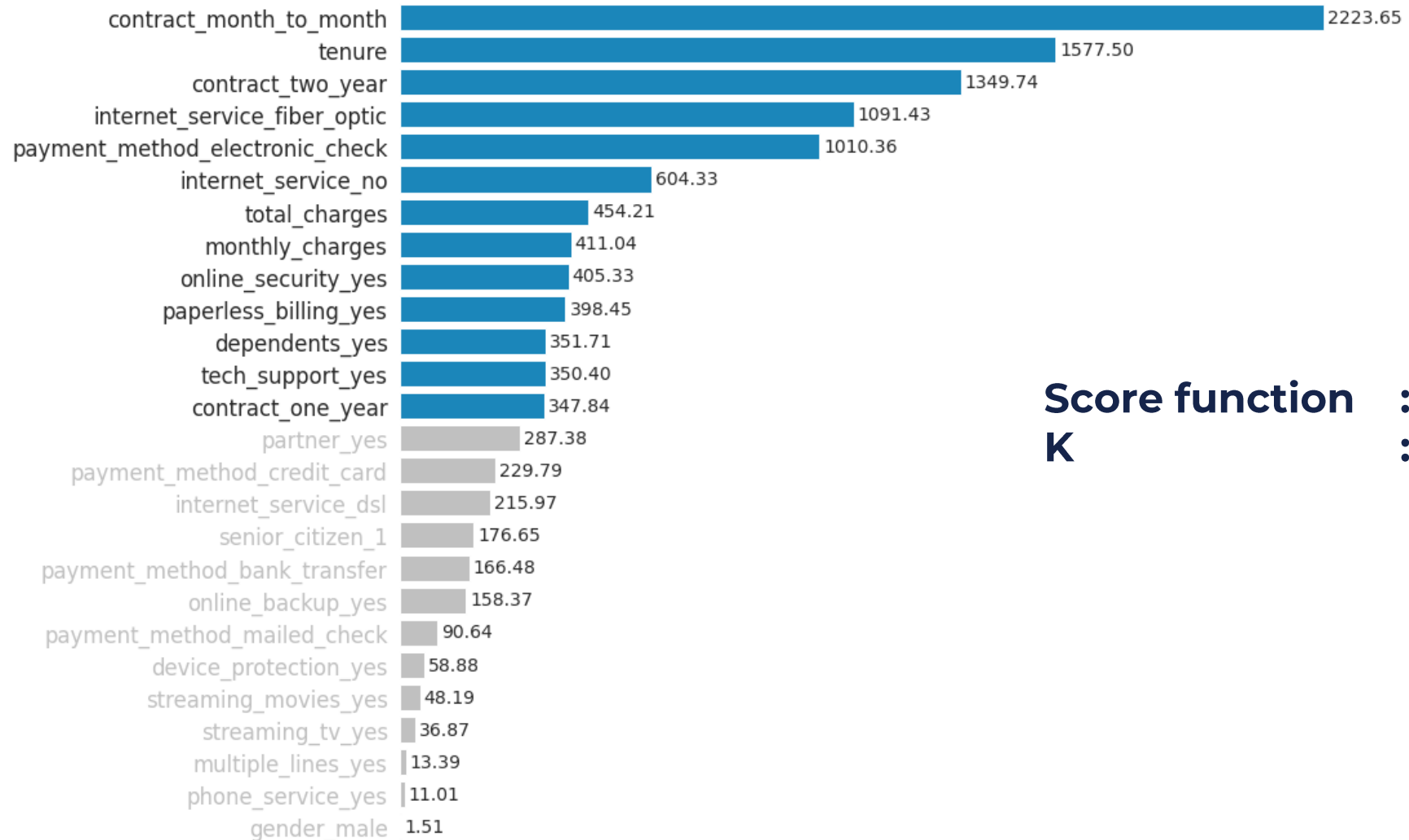
	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.747	0.707	0.755	0.714	0.755
Ridge Classifier	0.744	0.707	0.756	0.713	0.756
KNN	0.696	0.660	0.699	0.661	0.699
SVC	0.765	0.713	0.747	0.724	0.747
Neural Network	0.752	0.686	0.696	0.690	0.696
Decision Tree	0.731	0.666	0.682	0.672	0.682
Random Forest	0.771	0.708	0.714	0.711	0.714
<b>Gradient Boosting Classifier</b>	<b>0.788</b>	<b>0.734</b>	<b>0.763</b>	<b>0.744</b>	<b>0.763</b>
AdaBoost Classifier	0.755	0.712	0.756	0.720	0.756
CatBoost Classifier	0.786	0.725	0.728	0.727	0.728
Hist Gradient Boosting	0.780	0.719	0.722	0.721	0.722
XGBoost	0.784	0.731	0.762	0.741	0.762
LightGBM	0.785	0.725	0.732	0.728	0.732



Note:  
Precision, recall, F1 score, and ROC AUC score are calculated using **macro average**

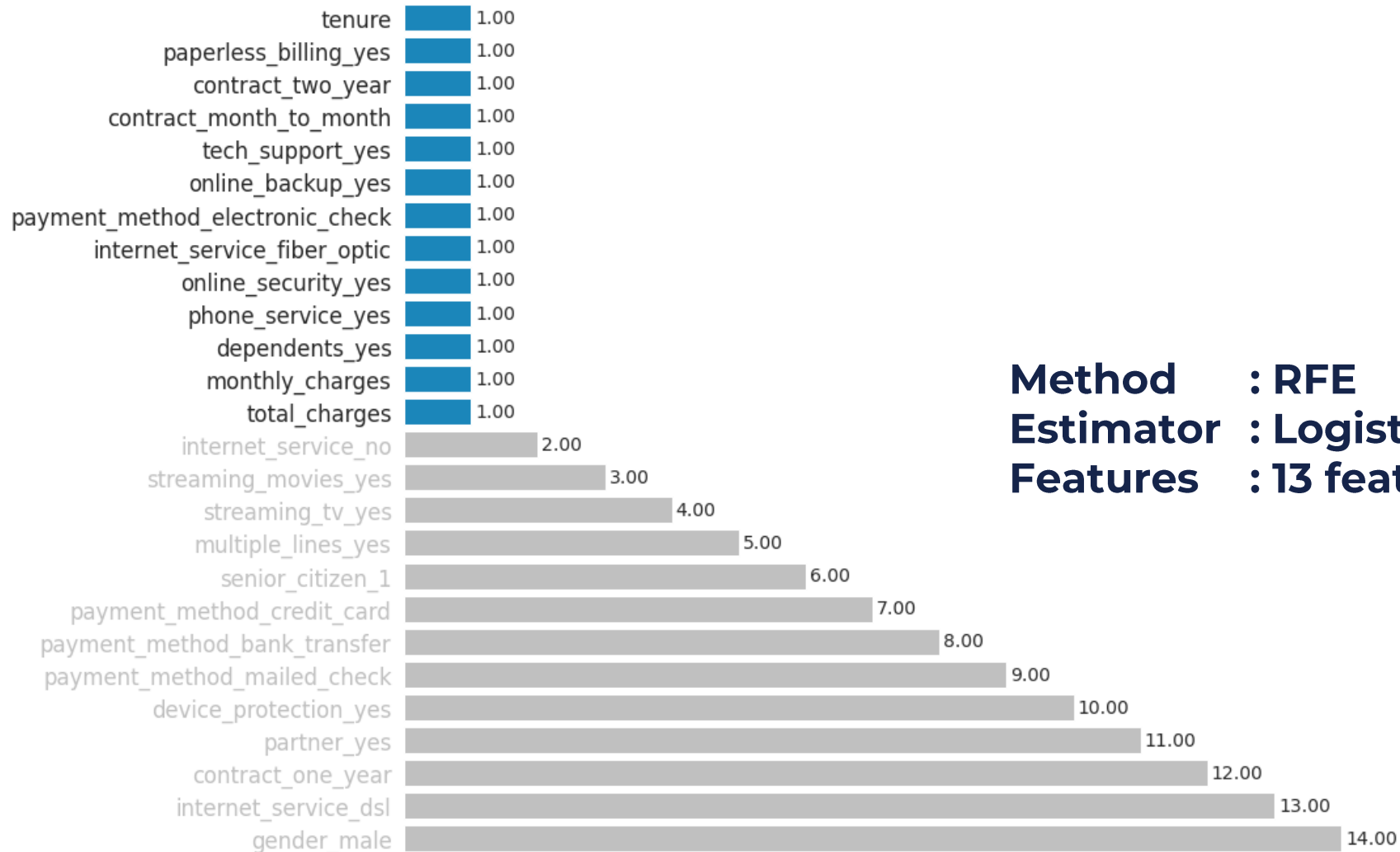
Overall, **boosting methods** show a good performance  
Can we improve it?

# Feature Selection – Filter Method



**Score function : ANOVA**  
**K : 13 features**

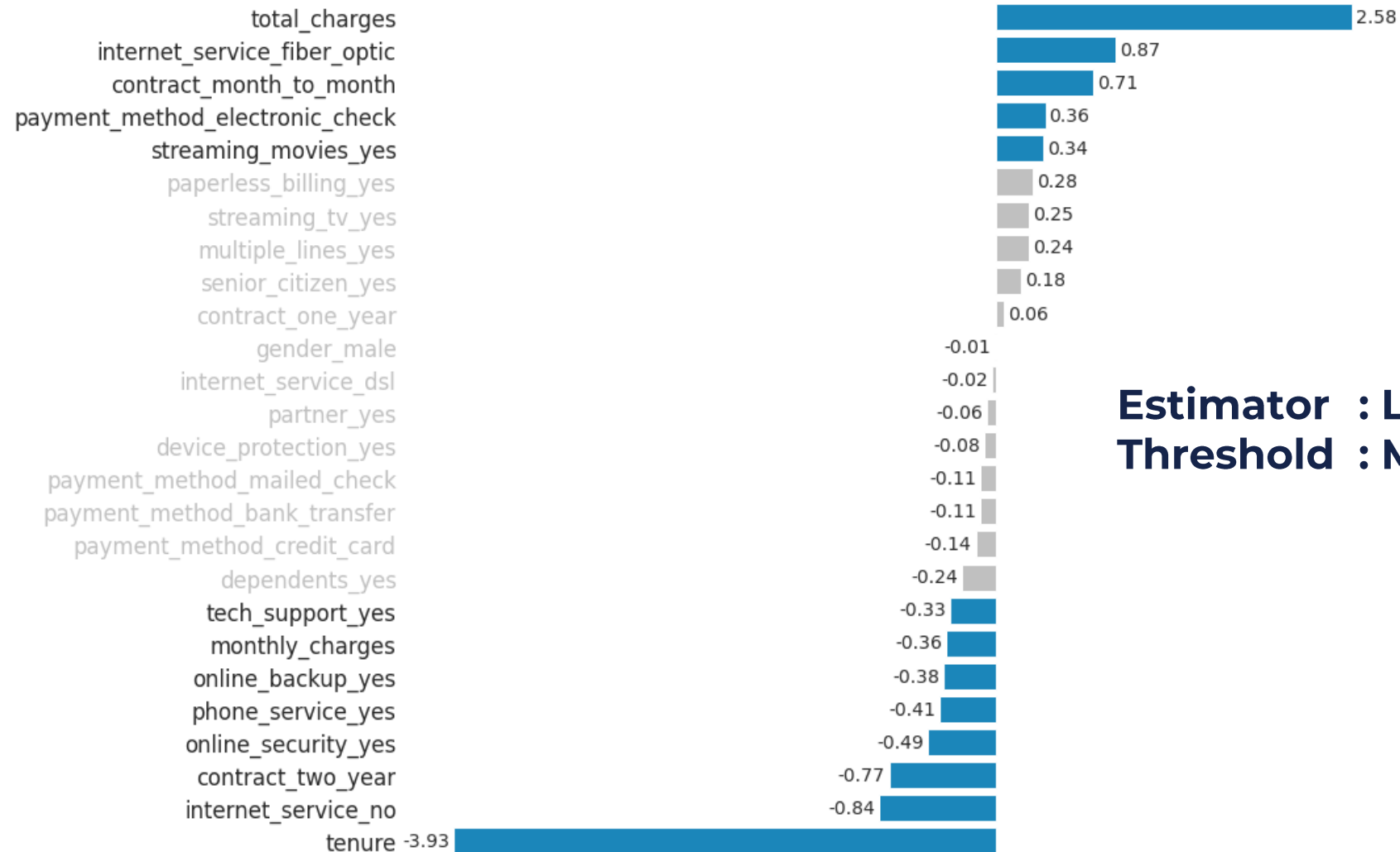
# Feature Selection – Wrapper Method



**Method** : RFE  
**Estimator** : Logistic Regression  
**Features** : 13 features



# Feature Selection – Embedded Method



**Estimator : Logistic Regression**  
**Threshold : Median**

# Feature Selection Comparison

The average metrics of boosting models using default parameter

	Accuracy	Precision	Recall	F1 Score	ROC AUC	
Without Feature Selection	0.780	0.724	0.744	0.730	0.744	 high    low
Filter Method	0.773	0.722	0.753	0.731	0.753	
Wrapper Method	0.775	0.722	0.751	0.731	0.751	
Embedded Method	0.766	0.714	0.746	0.723	0.746	

In general, the **filter method** shows the best result  
Moreover, it has the **highest recall** score

Note:  
Precision, recall, F1 score, and ROC AUC score are calculated using **macro average**

# Hyperparameter Tuning Strategy

## Adaboost classifier without feature selection


Default Parameter				Tuning 1				Tuning 2			
	precision	recall	f1-score		precision	recall	f1-score		precision	recall	f1-score
0	0.896	0.753	0.819	0	0.872	0.841	0.856	0	0.902	0.758	0.824
1	0.527	0.759	0.622	1	0.600	0.660	0.628	1	0.536	0.772	0.633
macro avg	0.712	0.756	0.720	macro avg	0.736	0.750	0.742	macro avg	0.719	0.765	0.728
weighted avg	0.798	0.755	0.766	weighted avg	0.800	0.793	0.796	weighted avg	0.805	0.762	0.773
accuracy		0.755		accuracy		0.793		accuracy		0.762	
roc auc		0.756		roc auc		0.750		roc auc		0.765	

I choose the second tuning because it has a higher positive recall than the first tuning

The hyperparameter tuning strategy is focused on  
Maximizing the **recall** score of the **positive class** (churn), not the average  
But still paying attention to the **accuracy** score

# Metrics after Tuning


## Without Feature Selection

	Accuracy		Recall (Positive)		
	Before	After	Before	After	
Gradient Boosting Classifier	0.788	0.783	0.709	0.742	 high low
AdaBoost Classifier	0.755	0.762	0.759	0.772	
CatBoost Classifier	0.786	0.772	0.606	0.740	
Hist Gradient Boosting	0.780	0.763	0.599	0.774	
XGBoost	0.784	0.772	0.715	0.763	
LightGBM	0.785	0.760	0.619	0.783	

After tuning, the accuracy score is mostly decreased  
But, the recall score has increased dramatically

# Metrics after Tuning

## With feature selection

	Accuracy		Recall (Positive)		
	Ori	Tuned	Ori	Tuned	
Gradient Boosting Classifier	0.772	0.775	0.725	0.766	 high low
AdaBoost Classifier	0.755	0.759	0.777	0.783	
CatBoost Classifier	0.788	0.761	0.677	0.765	
Hist Gradient Boosting	0.773	0.756	0.668	0.781	
XGBoost	0.774	0.761	0.752	0.779	
LightGBM	0.777	0.762	0.663	0.791	

After tuning, the accuracy score is mostly decreased  
But, the recall score has increased dramatically

# Model Selection

Using **harmonic mean** of the accuracy and recall scores

$$F_{\beta} = (1 + \beta^2) \frac{(\text{accuracy} * \text{recall})}{\beta^2 * \text{accuracy} + \text{recall}}$$

## Without feature selection

## With feature selection

	Accuracy	Recall (Positive)	F-beta	Accuracy	Recall (Positive)	F-beta
Gradient Boosting Classifier	0.783	0.742	0.762	0.775	0.766	0.770
AdaBoost Classifier	0.762	0.772	0.767	0.759	0.783	0.771
CatBoost Classifier	0.772	0.740	0.756	0.761	0.765	0.763
Hist Gradient Boosting	0.763	0.774	0.768	0.756	0.781	0.768
XGBoost	0.772	0.763	0.767	0.761	0.779	0.770
LightGBM	0.760	0.783	0.771	0.762	0.791	0.776

# Conclusion

## Final Model

- LightGBM with feature selection using filter method

## Recommendation and Request

- We should **pay more attention** to customers who meet the criteria below
  - Contract : Month-to-month
  - Tenure : Short tenure
  - Internet service : Fiber optic
  - Payment method : Electronic check
- Please, **evaluate our service!**  
Especially for internet service (fiber optic) and payment method (electronic check)
- Can we **give more benefit** to a new customer?  
Because the new customer has a high probability to churn



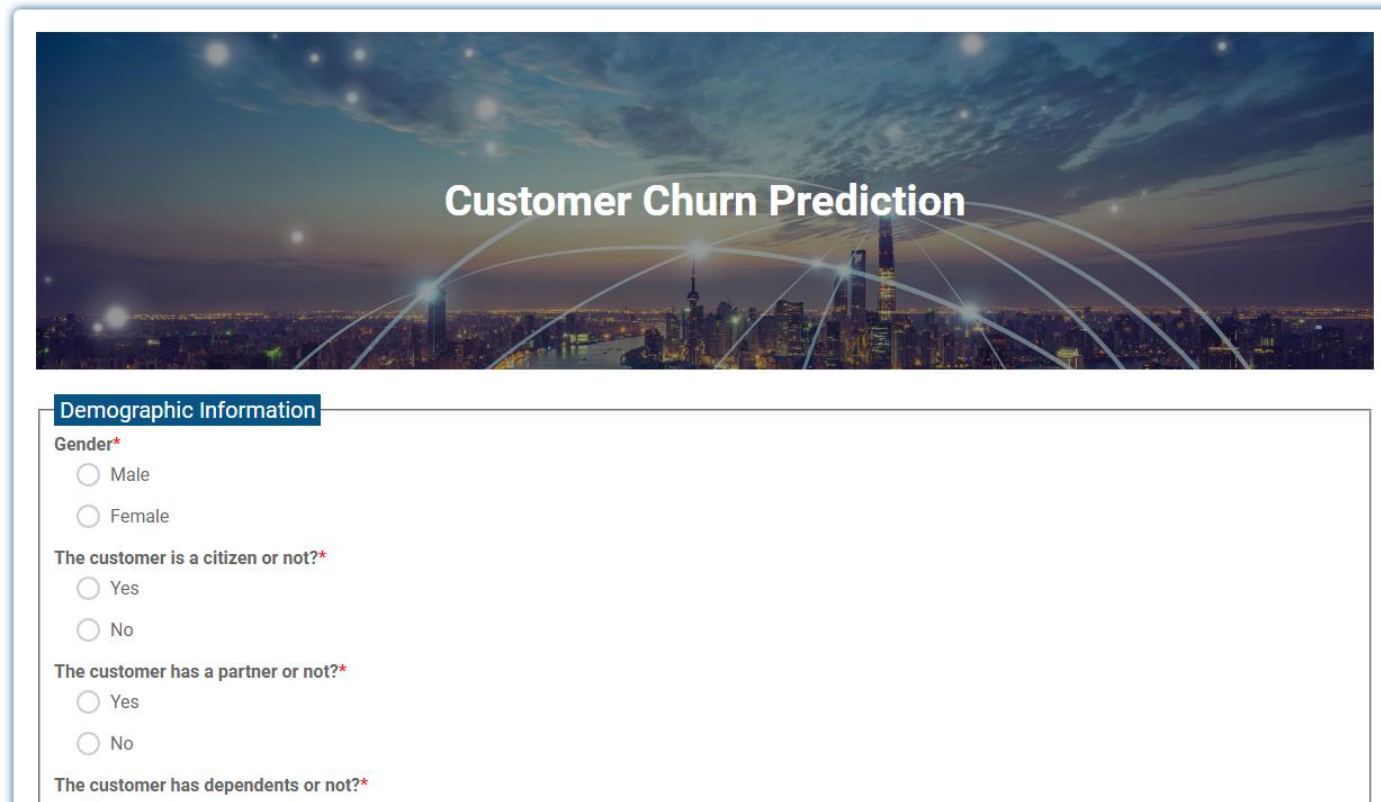
06

**BONUS!**

Last but not least



# Model Deployment



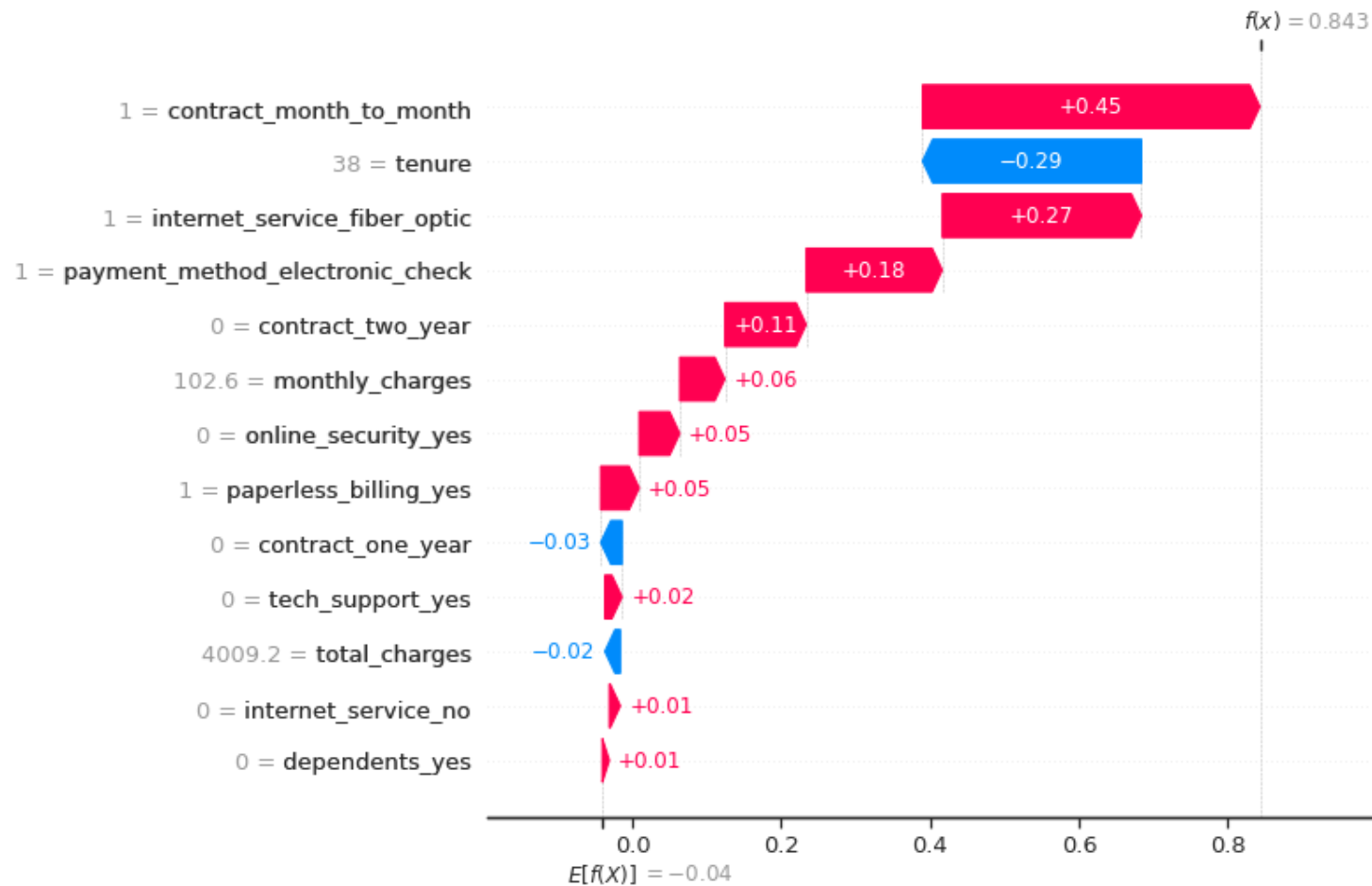
The screenshot displays a web application interface for 'Customer Churn Prediction'. The header features a cityscape at night with glowing network lines. Below the header is a form titled 'Demographic Information' with the following fields:

- Gender\***
  - Male
  - Female
- The customer is a citizen or not?\***
  - Yes
  - No
- The customer has a partner or not?\***
  - Yes
  - No
- The customer has dependents or not?\***
  - Yes
  - No

Model deployment using **Flask** and **Heroku**

<https://adhang-churn.herokuapp.com/>

# SHAP Explainable AI



SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model.

I have included it on my Heroku app

# THANKS

---

Adhang Muntaha Muhammad

CREDITS: This presentation template was originally created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**