# CUSTOMERS CHURN

SOUMIA ZARKAN - VICTOR GRAU -
DAVID ROJAS - MARIE LE CHEVALIER

# Business Context & Problem

## BUSINESS CONTEXT

With so many telecom operators in the market, it is not easy to keep up with competitors in the sector or to understand what makes consumers leave for the competition. So, the telecommunications sector is subject to a high churn rate that needs to be analysed.

## BUSINESS PROBLEM

What are the factors of churn in the telecommunications sector, and in particular at Orange? What actions can be taken to reduce it?

# Our customers

### Customers who have internet

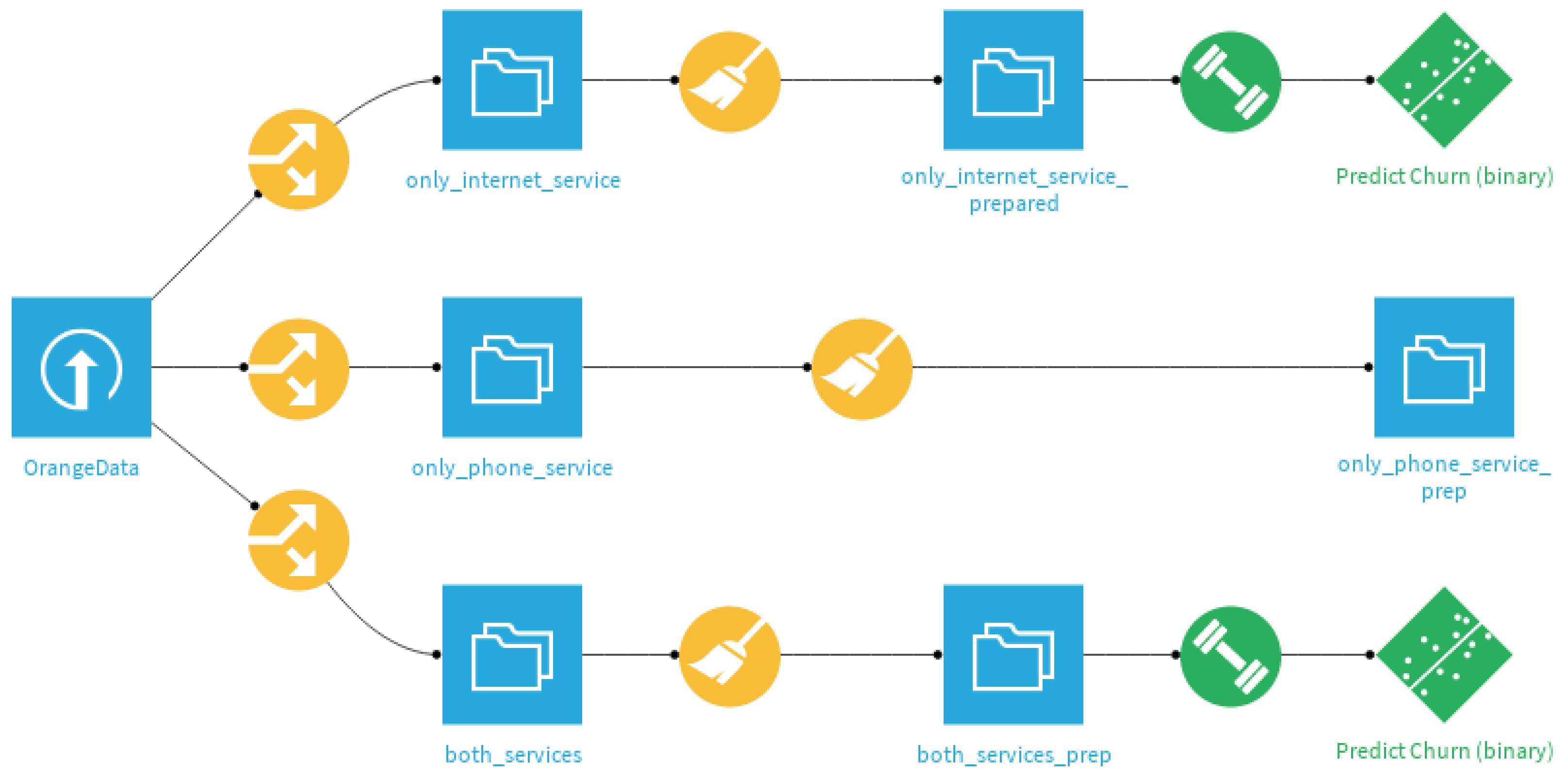
People with a subscription to the Orange operator for an internet box.

### Customers who have phone line(s) AND Internet

People with both a telephone line and an internet box with the Orange operator.

### Customers who don't have a phone line

People who do not have a telephone line with the Orange operator




First we decides to split our dataset in three distinct parts:

1. Customers with only phone service
2. Customers with only internet service
3. Customers with both services

We made the assumption that these three segments should be treated independently of each other.

After that we cleaned up the dataset (correcting inconsistencies, correcting types and deleting useless columns).

For the datasets with the internet service, we have also enriched the dataset with an additional column corresponding to the number of additional services (online security, tech support, streaming tv, streaming movies, device protection and online backup).

We did not make specific encodings on the categorical variables nor did we perform a reduced centered normalization on the quantitative varibels of the initial dataset since dataiku allows to manage these transformations at the time of the training of the various models.

Concerning the encoding of the variables, we noticed during the training of our models, that the target and the frequency encoding often allowed to obtain better performances.

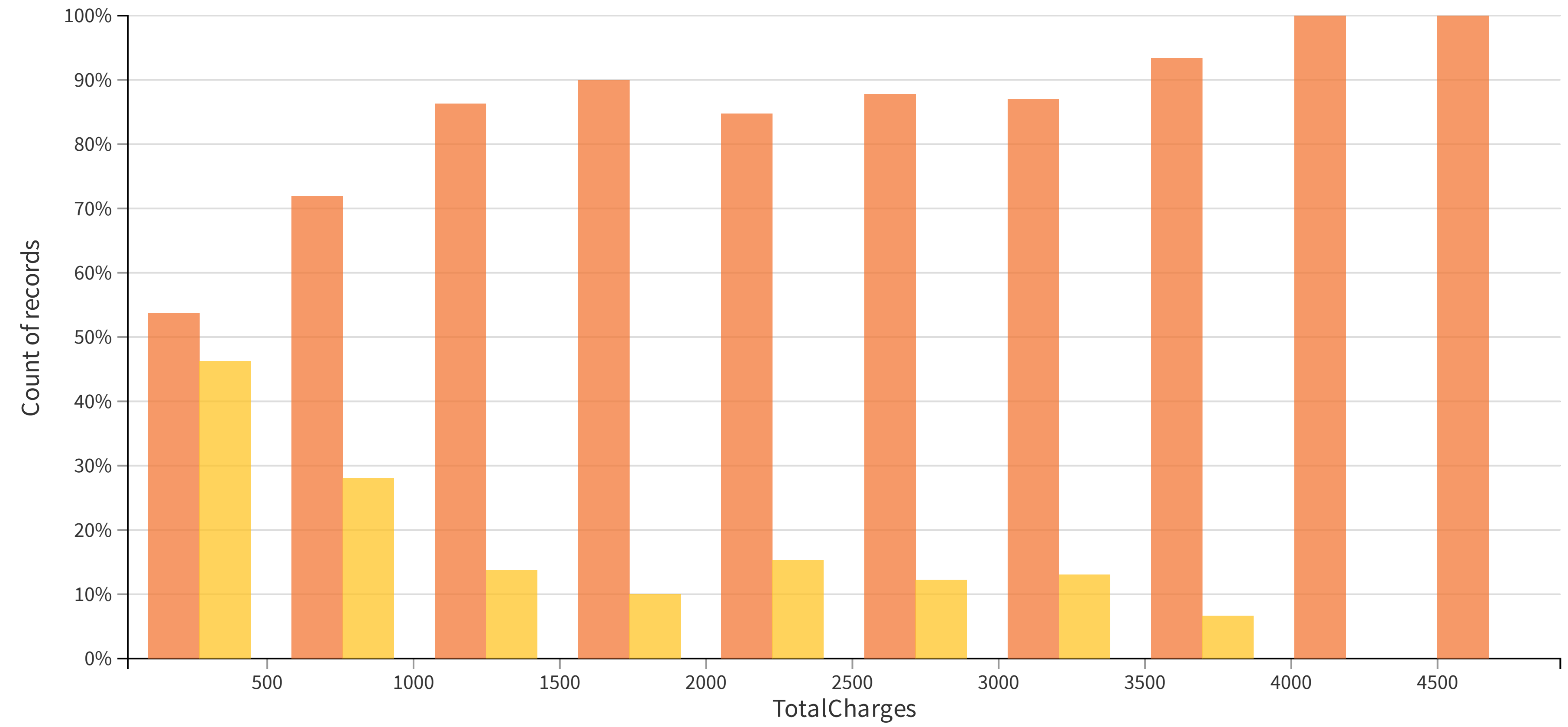
Data exploration

Only internet customers

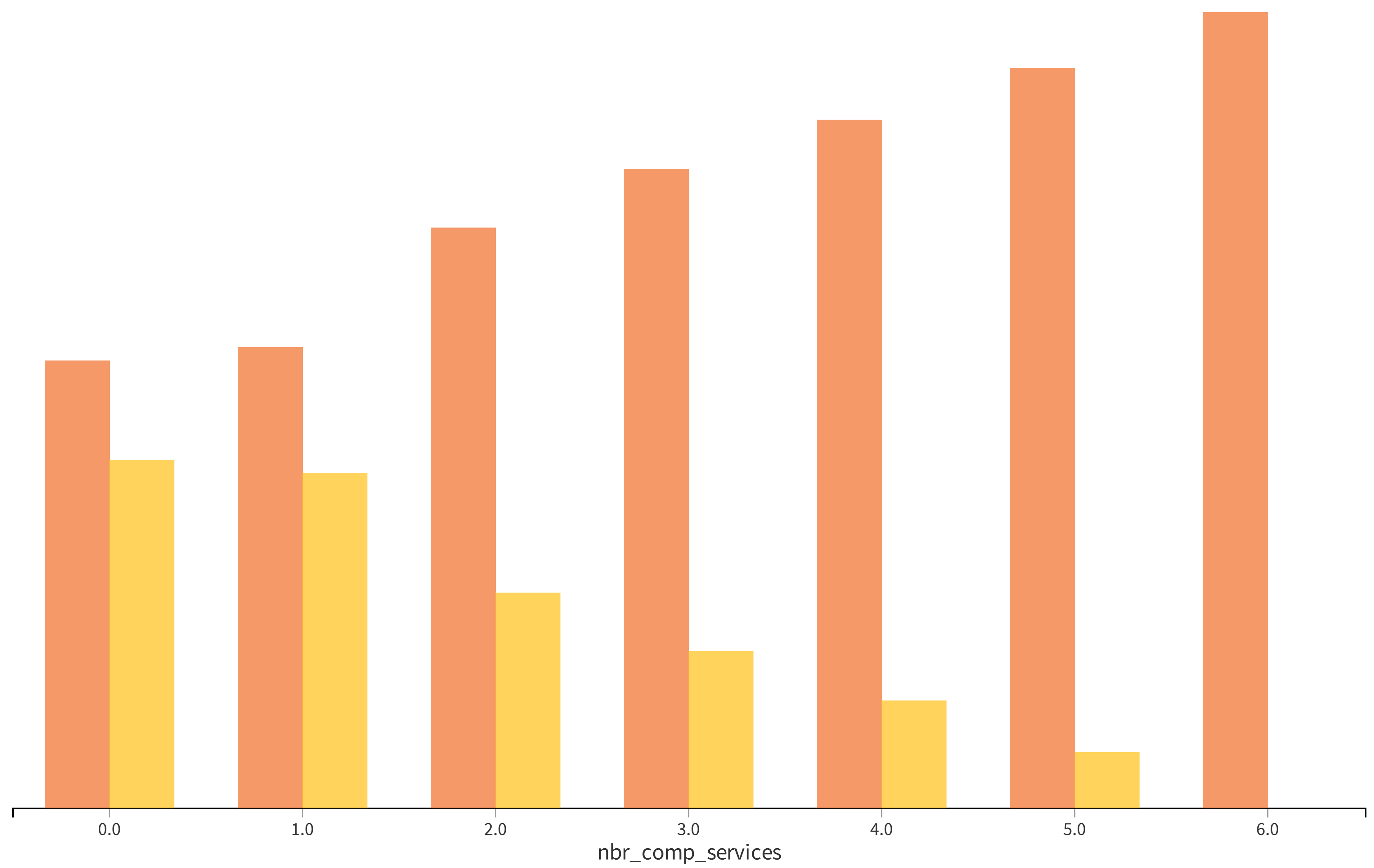
**Only internet service dataset**

| customerID | gender | SeniorCitizen | Partner | Dependents | tenure | OnlineSecurity | OnlineBackup | DevicePr |
|---|---|---|---|---|---|---|---|---|
| 7590-VHVEG | Female | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| 7795-CFOCW | Male | 0 | 0 | 0 | 45 | 1 | 0 | 1 |
| 6713-OKOMC | Female | 0 | 0 | 0 | 10 | 1 | 0 | 0 |
| 8779-QRDMV | Male | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| 8665-UTDHZ | Male | 0 | 1 | 1 | 1 | 0 | 1 | 0 |
| 0526-SXDJP | Male | 0 | 1 | 0 | 72 | 1 | 1 | 1 |
| 8108-UXRQN | Female | 0 | 1 | 1 | 11 | 1 | 0 | 0 |
| 3016-KSVCP | Male | 0 | 1 | 0 | 29 | 0 | 0 | 0 |
| 5386-THSLQ | Female | 1 | 1 | 0 | 66 | 0 | 1 | 1 |
| 6180-YBIQI | Male | 0 | 0 | 0 | 5 | 0 | 0 | 0 |
| 9750-BOOHV | Female | 0 | 0 | 0 | 32 | 1 | 0 | 0 |
| 5256-SKJGO | Female | 0 | 1 | 1 | 64 | 0 | 1 | 0 |
| 9560-BBZXK | Female | 0 | 0 | 0 | 36 | 1 | 0 | 0 |
| 2639-UGMAZ | Male | 1 | 0 | 0 | 71 | 1 | 1 | 0 |

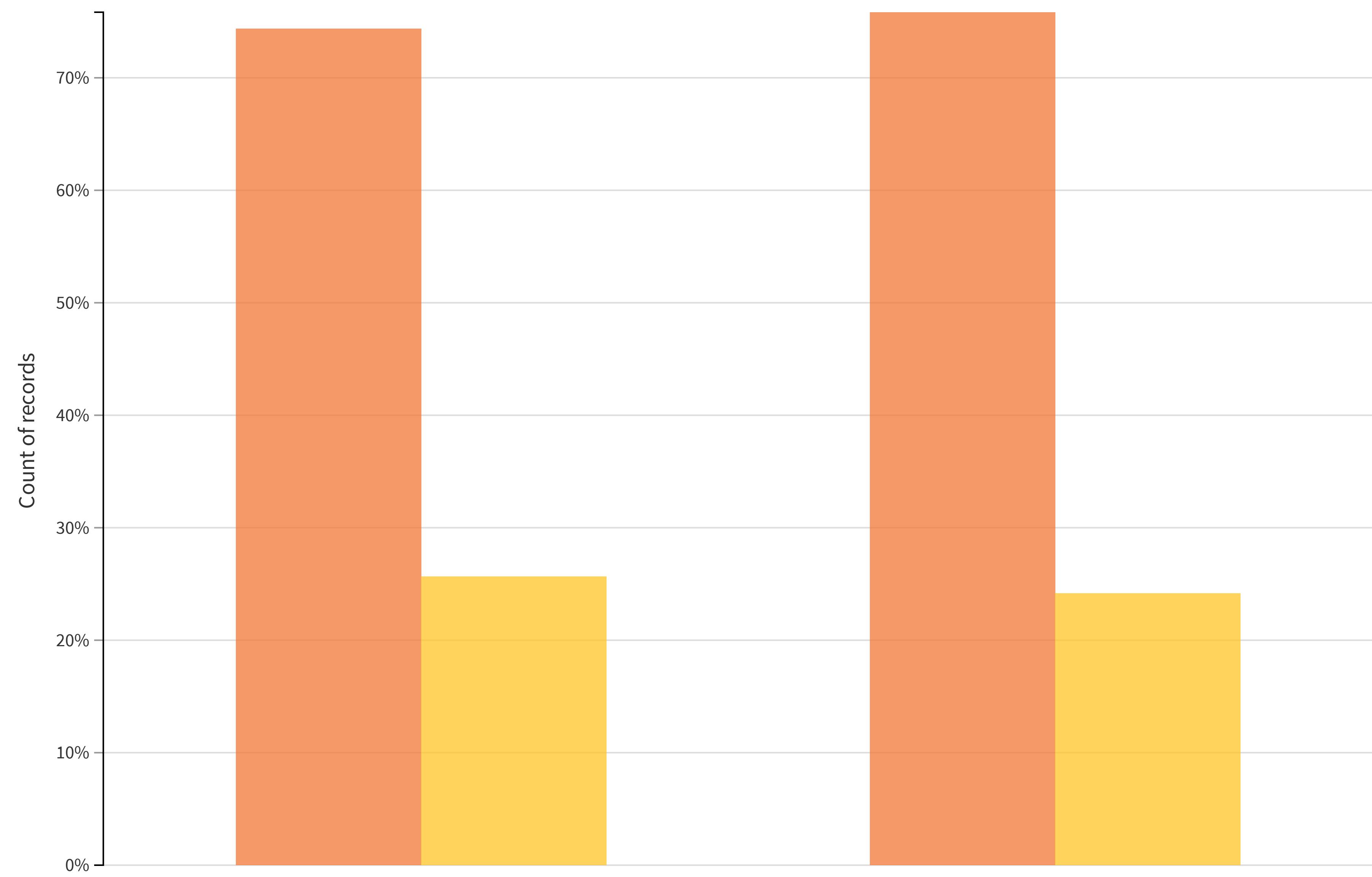
**Total Charges impact on churn**




**Number of services subscribed and impact on churn**




**Impact of gender on churn**



Correlation matrix on 3 variables (Pearson)

| | tenure | MonthlyCharges | TotalCharges |
|---|---|---|---|
| tenure | 1.000 | 0.586 | 0.953 |
| MonthlyCha... | 0.586 | 1.000 | 0.747 |
| TotalCharges | 0.953 | 0.747 | 1.000 |

Observations

- We checked the correlation between the inputs variables and the target variable. We saw with different graphics that some variables are not correlated with the churn like the gender.
- We also checked the correlation between numerical input variables, and logically saw a strong correlation between the tenure and total charges. Most of information contains in the tenure is also contained in total charges (seniority). So we can make the hypothesis that the tenure is redundant regarding total charges.

Only internet service ML analysis

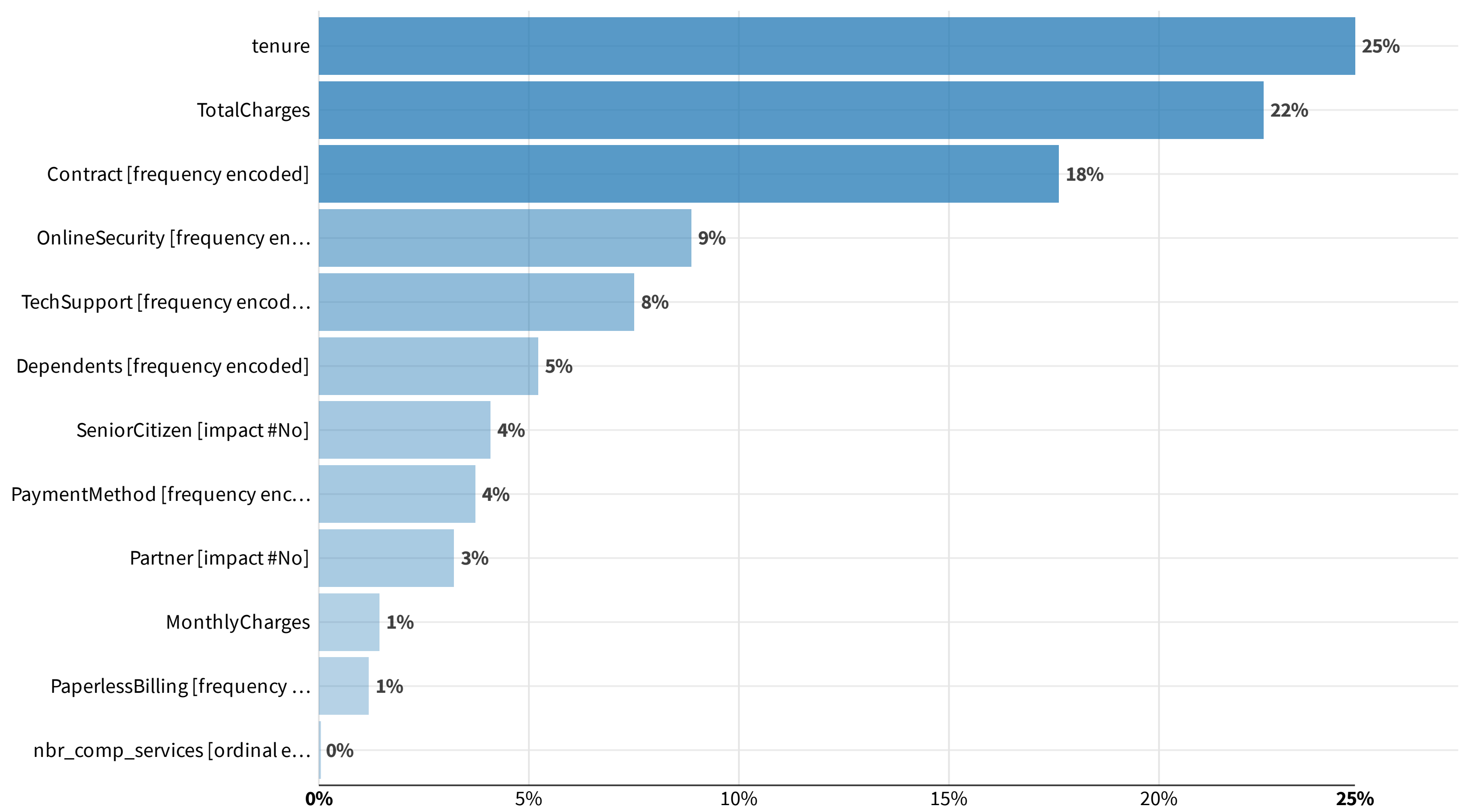
# LightGBM (s42) - v2

*ROC AUC: 0.830*

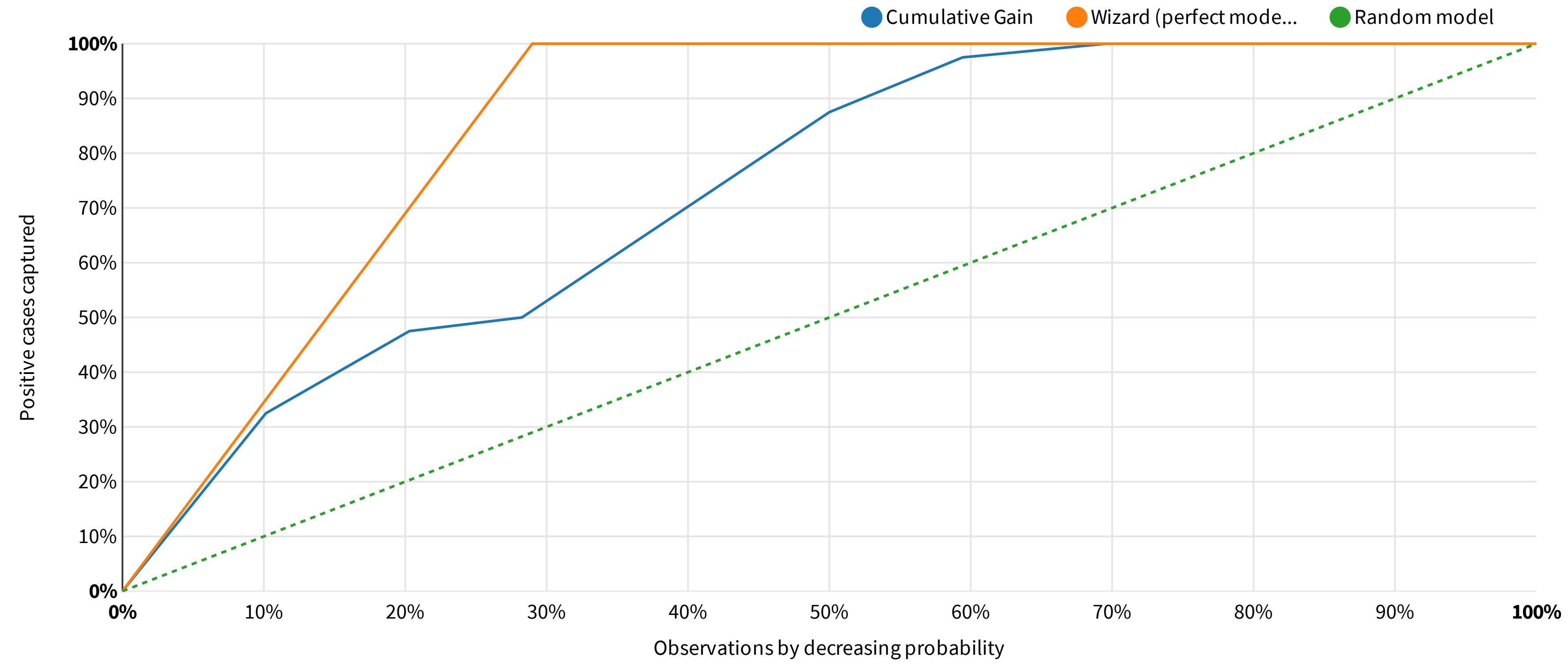
## Model

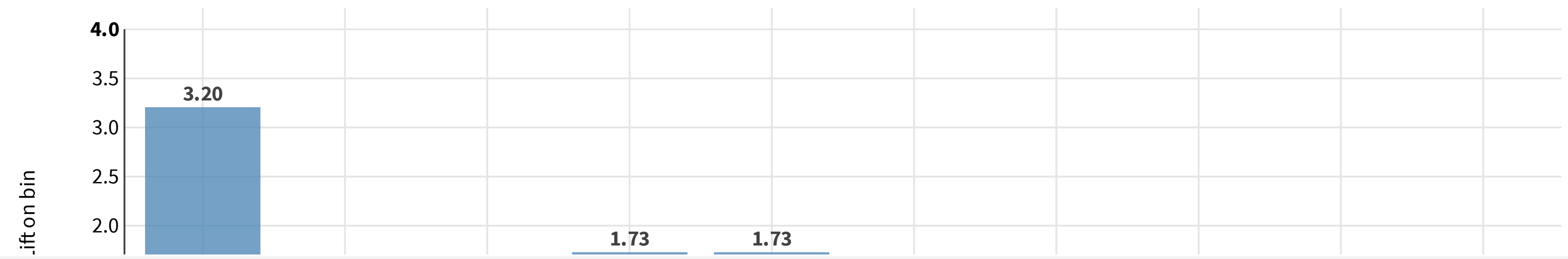
| | |
|---|---|
| Model ID | **S-ML8-yjeVY26T-1671098215571** |
| Backend | **Python (in memory)** |
| Algorithm | **Lightgbm classification** |
| Trained on | **2022/12/15 10:56** |

## Variable importance




## Lift charts



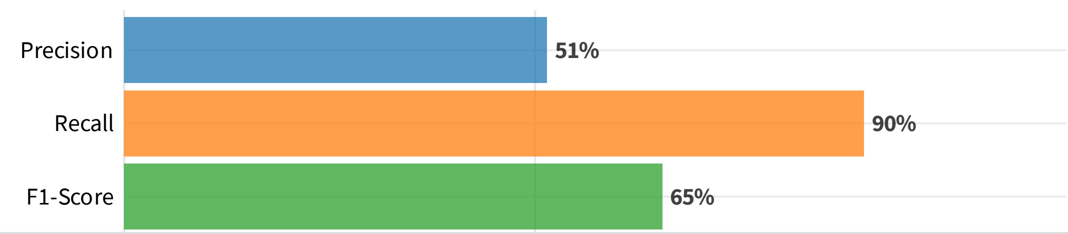



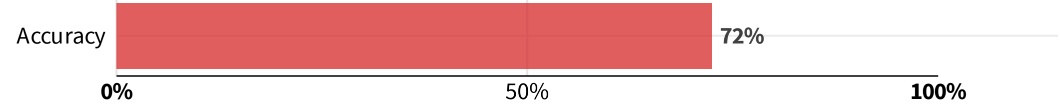

## Confusion matrix

Threshold (cut-off) 0 — 1 0.400 BACK TO OPTIMAL*

Display: Record count

| | Predicted Yes | Predicted No | Total |
|---|---|---|---|
| Actually Yes | 36 | 4 | 40 |
| Actually No | 34 | 64 | 98 |
| Total | 70 | 68 | 138 |





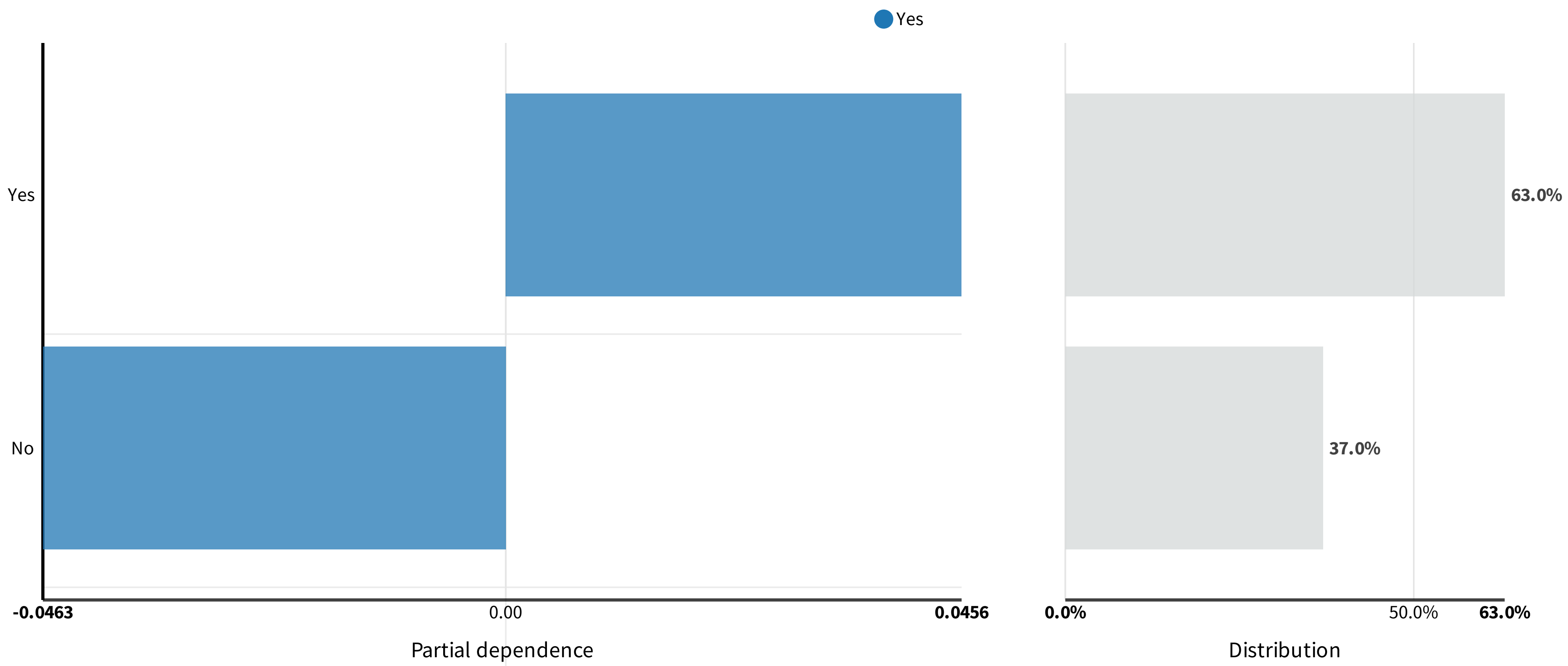

## Cost matrix

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| If model predicts Yes | and value is Yes | the gain is | 1 | × | 36 | = | 36.00 |
| | but value is No | the gain is | -0.3 | × | 34 | = | -10.20 |
| Model predicts No | and value is No | the gain is | 0 | × | 64 | = | 0.00 |
| | but value is Yes | the gain is | 0 | × | 4 | = | 0.00 |
| | **Average gain per record** | | 0.19 | × | 138 | = | 25.80 |

## Partial dependence

Select your variable A PaperlessBilling COMPUTE COMPUTE ALL EXPORT

2 most frequent modalities of **PaperlessBilling**, computed on 138 rows (the full test set)



As we can see in the left figures, the two most important variables in our model (LightGBM) are the tenure and the total charges, which are highly correlated. we also have the type of contract in third position. Concerning the complementary services, the security and tech support services seem to be the more important for the customers whereas other complementary services don't seem so much important.

By analysing the partial dependence of the variables with the churn, we can make some observations:

- For the total charges, we detect a threshold value at 900. Above this value, the customers are more likely to stay
- For the tenure, above 24 months (2 years), the customers are also more likely to stay correlated with its seniority
- For the contract type, the customer has more probability to leave with a month-to-month contract while a one year customer has more probability to stay. This probability is even stronger with a 2-year contract customer (link with the threshold detected in the tenure dependence)

# Our Recommendations

## (Internet service)

General target characteristics

1. Improve and focus marketing campaigns on customers with low seniority. For example, offers on complementary services as tech support and security
2. If possible, limit the sale of contracts month by month
3. Focus on family offers as much on possible
4. Focus marketing campaigns on senior citizens
5. Focucs marketing campaigns on people alone
6. Focus marketing campaigns on people paying by electronic checks

Now if you want to predict if a specific customer is going to churn or not, you can run our LightGBM model with the target customer inputs. Depending on the results and the marketing budget, the following 3 strategies are recommended to target churn-prone customers
If you want a good trade off between false negative and positive, you can set the threshold at 0.4. In this case, we can estimate that 50% of the customers with only internet service will churn soon. On these predictions, 50% are really about to churn.  On the 50 other percents that are not detected for churn, 6% are really about to churn.
If you don't want to miss churn, you can fix the threshold to 0.3. In this case, you will have to target 67% of your customers, and of those, 43% will be susceptible to churns.
For a limited budget, and to be sure to target customers really likely to churn, the threshold should be set at least at 0.725. This represents 9% of the total customer base that only has internet service.  We can estimate that among the remaining 90% of customers, 22% are susceptible to churn.

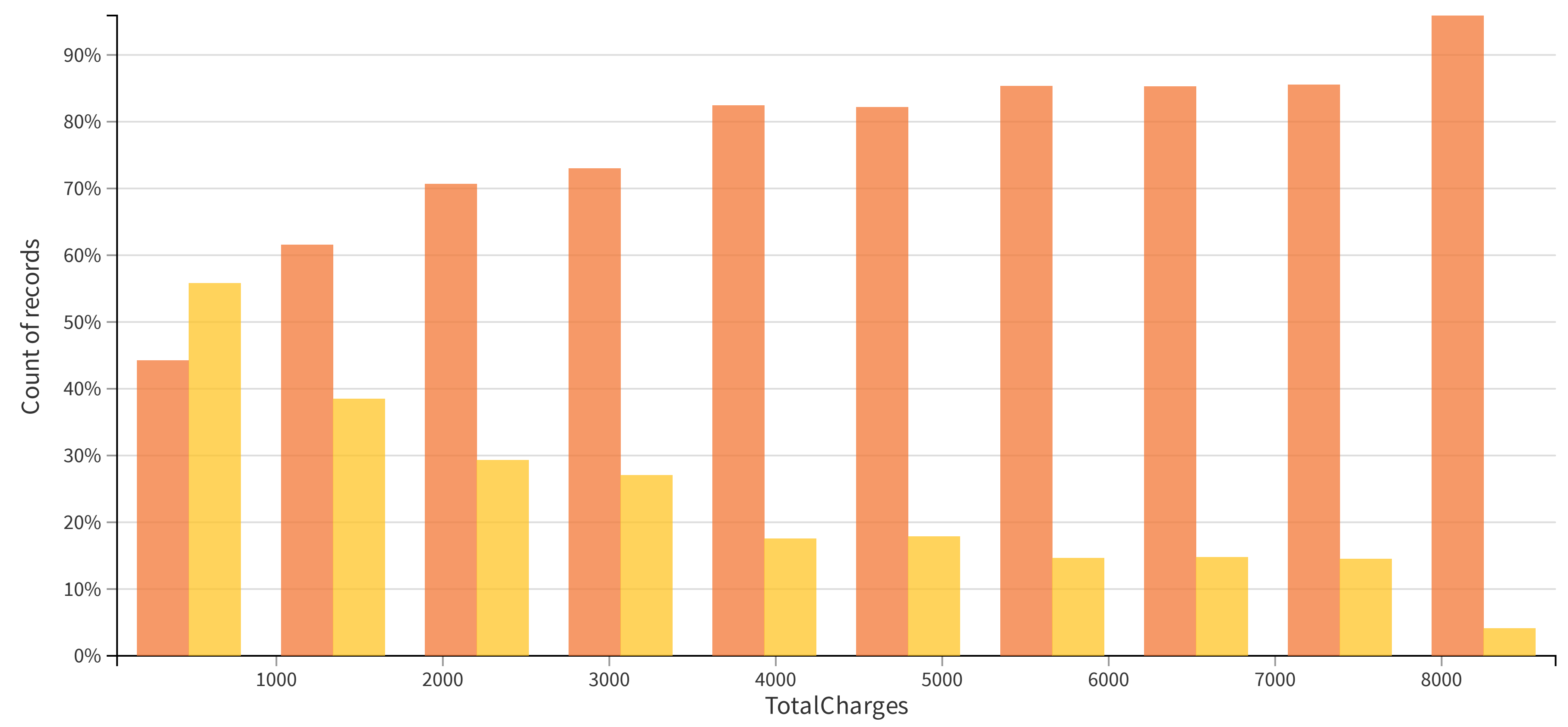
Data exploration

Both Services customers

## Both services dataset customers

| customerID | gender | SeniorCitizen | Partner | Dependents | tenure | PhoneService | MultipleLines | InternetS |
|---|---|---|---|---|---|---|---|---|
| 5575-GNVDE | Male | 0 | 0 | 0 | 34 | Yes | 0 | DSL |
| 3668-QPYBK | Male | 0 | 0 | 0 | 2 | Yes | 0 | DSL |
| 9237-HQITU | Female | 0 | 0 | 0 | 2 | Yes | 0 | Fiber optic |
| 9305-CDSKC | Female | 0 | 0 | 0 | 8 | Yes | 1 | Fiber optic |
| 1452-KIOVK | Male | 0 | 0 | 1 | 22 | Yes | 1 | Fiber optic |
| 7892-POOKP | Female | 0 | 1 | 0 | 28 | Yes | 1 | Fiber optic |
| 6388-TABGU | Male | 0 | 0 | 1 | 62 | Yes | 0 | DSL |
| 9763-GRSKD | Male | 0 | 1 | 1 | 13 | Yes | 0 | DSL |
| 8091-TTVAX | Male | 0 | 1 | 0 | 58 | Yes | 1 | Fiber optic |
| 0280-XJGEX | Male | 0 | 0 | 0 | 49 | Yes | 1 | Fiber optic |
| 5129-JLPIS | Male | 0 | 0 | 0 | 25 | Yes | 0 | Fiber optic |
| 3655-SNQYZ | Female | 0 | 1 | 1 | 69 | Yes | 1 | Fiber optic |
| 9959-WOFKT | Male | 0 | 0 | 1 | 71 | Yes | 1 | Fiber optic |
| 4190-MFLUW | Female | 0 | 1 | 1 | 10 | Yes | 0 | DSL |
| 4183-MYFRB | Female | 0 | 0 | 0 | 21 | Yes | 0 | Fiber optic |
| 3638-WEABW | Female | 0 | 1 | 0 | 58 | Yes | 1 | DSL |
| 6322-HRPFA | Male | 0 | 1 | 1 | 49 | Yes | 0 | DSL |
| 6865-JZNKO | Female | 0 | 0 | 0 | 30 | Yes | 0 | DSL |
| 6467-CHFZW | Male | 0 | 1 | 1 | 47 | Yes | 1 | Fiber optic |
| 5248-YGIJN | Male | 0 | 1 | 0 | 72 | Yes | 1 | DSL |
| 8773-HHUOZ | Female | 0 | 0 | 1 | 17 | Yes | 0 | DSL |
| 3841-NFECX | Female | 1 | 1 | 0 | 71 | Yes | 1 | Fiber optic |
| 4929-XIHVW | Male | 1 | 1 | 0 | 2 | Yes | 0 | Fiber optic |
| 6827-IEAUQ | Female | 0 | 1 | 1 | 27 | Yes | 0 | DSL |
| 3413-BMNZE | Male | 1 | 0 | 0 | 1 | Yes | 0 | DSL |
| 6234-RAAPL | Female | 0 | 1 | 1 | 72 | Yes | 1 | Fiber optic |

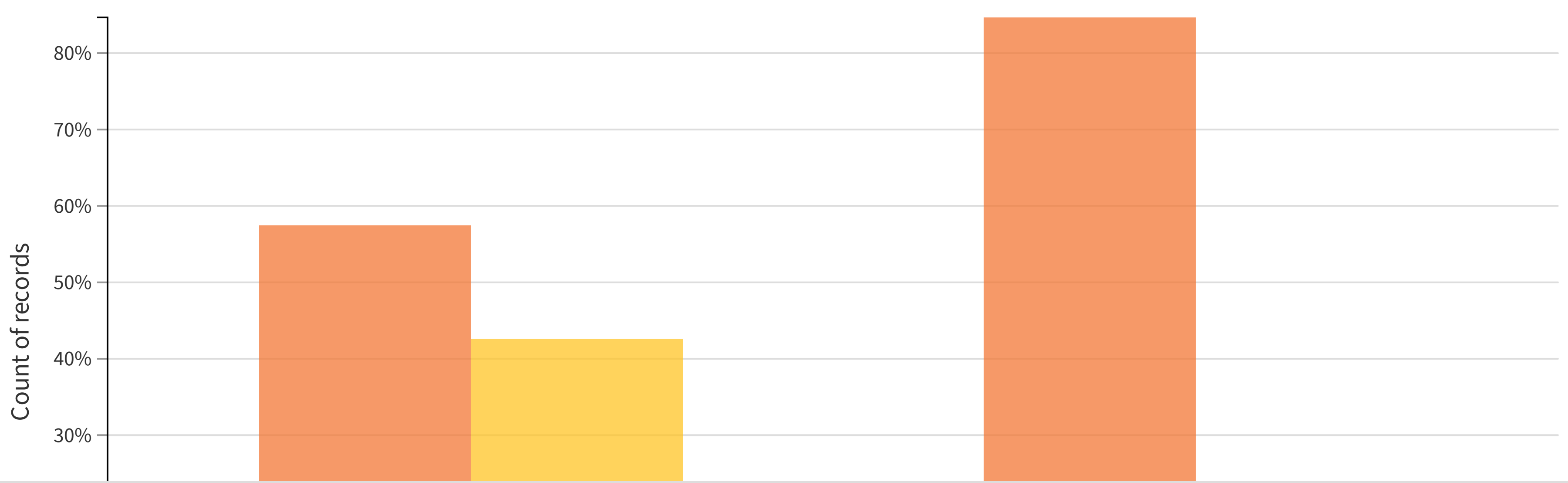
## Total Charges and Churn



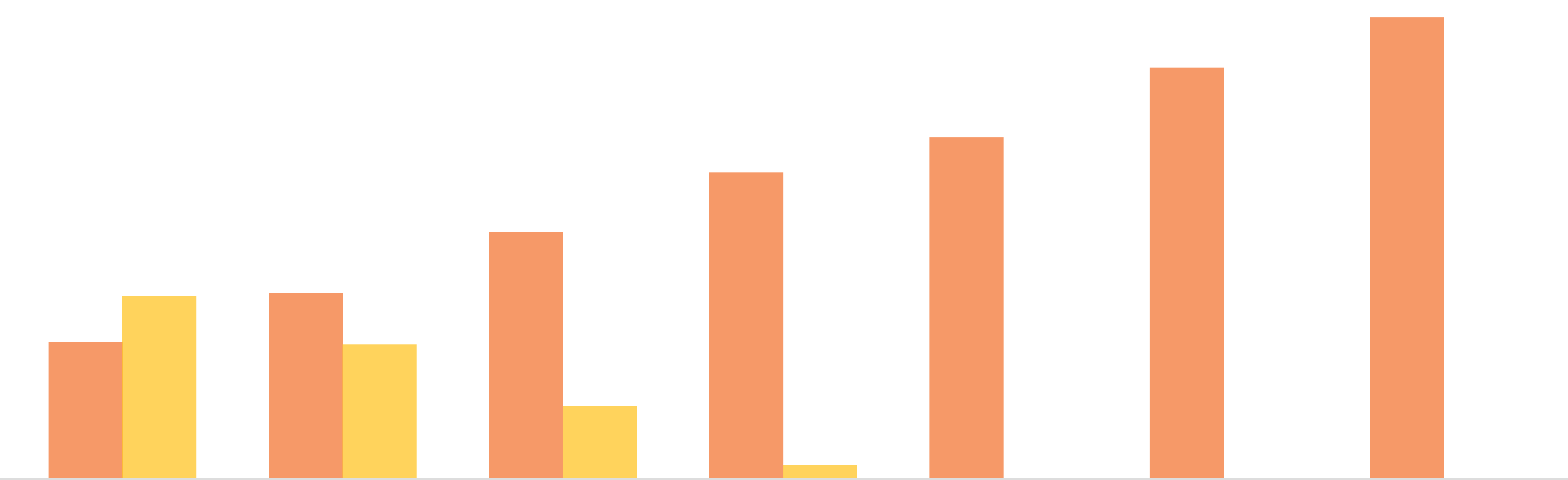

## Correlation matrix on 4 variables (Pearson)

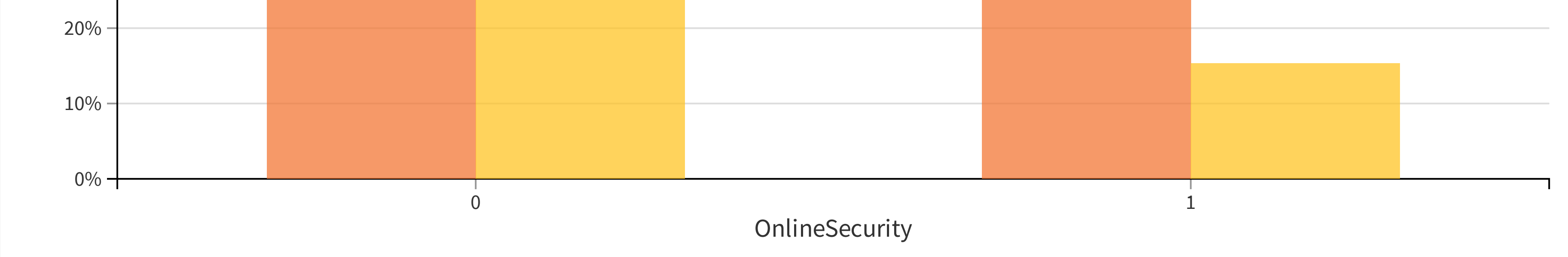
| | nb_multiple_services | TotalCharges | tenure | MonthlyCharges |
|---|---|---|---|---|
| nb_multipl... | 1.000 | 0.721 | 0.658 | 0.641 |
| TotalCharges | 0.721 | 1.000 | 0.958 | 0.623 |
| tenure | 0.658 | 0.958 | 1.000 | 0.443 |
| MonthlyCha... | 0.641 | 0.623 | 0.443 | 1.000 |

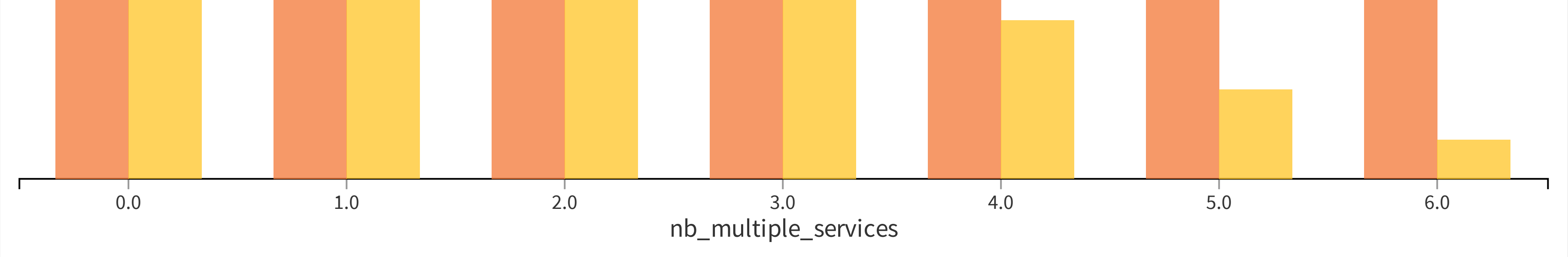
## Online security impact on churn




## Number of services subscribed impact on churn





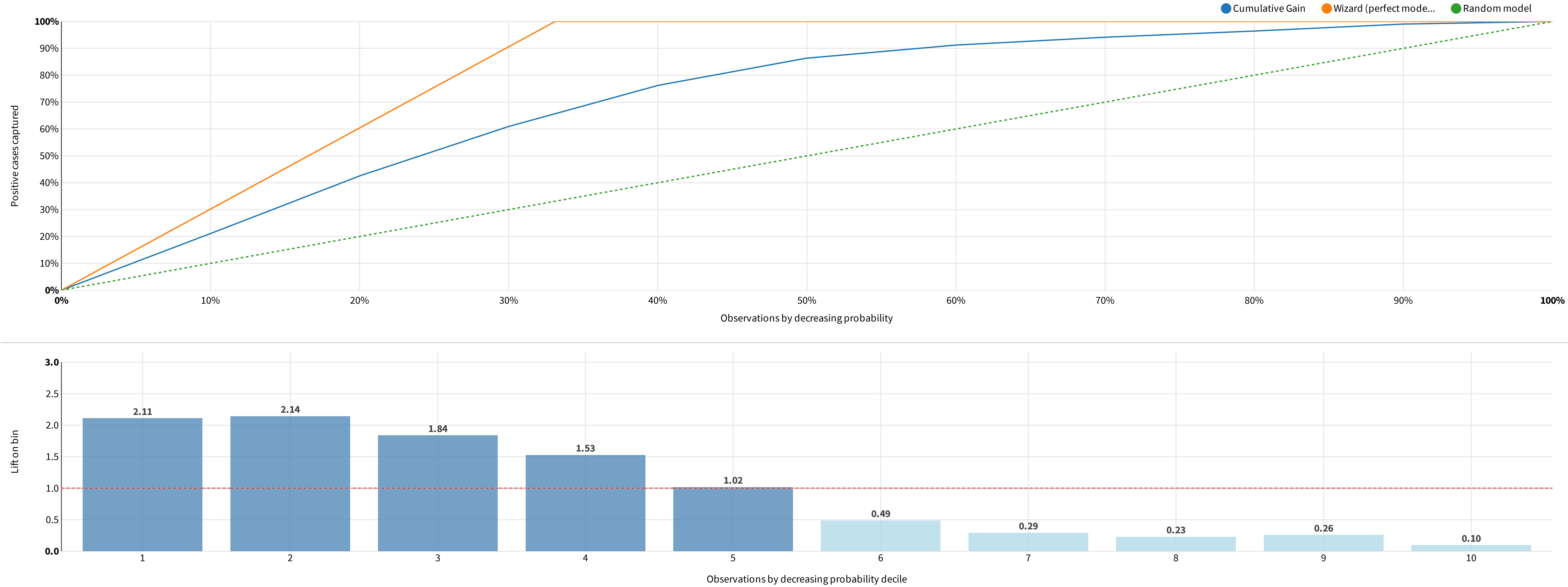




**Observations**

- With histograms we checked the correlation between the inputs variables and the target variable. We can see that some variables might have an impact on churn, like the number of services, online security and total charges.
- We can see that the proportion of churn is higher than the proportion of non churn when the number of complementary services is equal to 0.

Both service ML analysis

## Lift charts




## Confusion matrix

Threshold (cut-off) 0 1 0.500 BACK TO OPTIMAL*

Display: Record count

| | Predicted Yes | Predicted No | Total |
|---|---|---|---|
| Actually Yes | 254 | 53 | 307 |
| Actually No | 164 | 456 | 620 |
| Total | 418 | 509 | 927 |

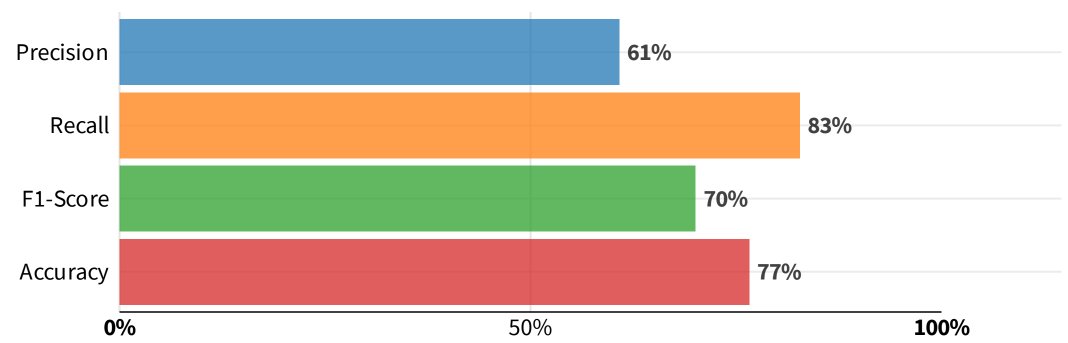
## Logistic Regression (model optimisé AUC) - v1

*ROC AUC: 0.830*

### Model

| | |
|---|---|
| Model ID | **S-ML8-5q2eMxSh-initial** |
| Backend | **Python (in memory)** |
| Algorithm | **Logistic regression** |
| Trained on | **2022/12/14 16:55** |

| | |
|---|---|
| Columns | 21 |
| Train set rows | 3908 |
| Test set rows | 927 |
| Calibration method | No calibration |



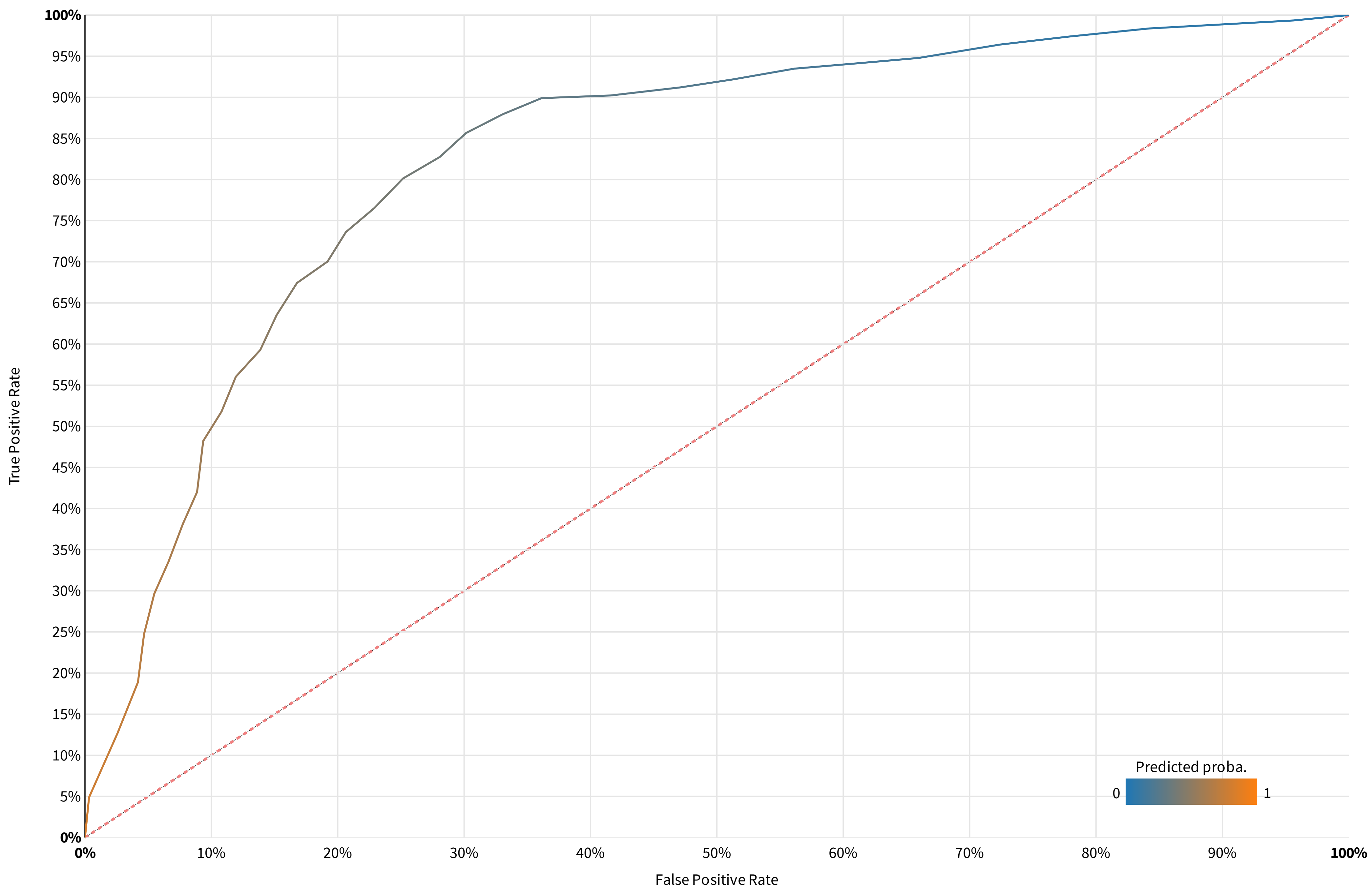

## Cost matrix

If model predicts and value is Yes the gain is 1 × 254 = 254.00

## Metadata

| | | |
|---|---|---|
| trainDataset:dataset-name | → | both_services_prep |
| testDataset:dataset-name | → | both_services_prep |
| evaluationDataset:dataset-name | → | both_services_prep |

## ROC curve




## Regression coefficients

EXPORT

Sort: | Coefficient | Filter Display coefficients for the unscaled variables

| Variable | Coefficient |
|---|---|
| `Contract` is `Month-to-month` | 1.2349 |
| `Contract` is `One year` | 0.6147 |
| `OnlineSecurity` is `0` | 0.5762 |
| `TechSupport` | -0.4677 |
| `PaperlessBilling` | 0.3732 |
| `OnlineBackup` is `0` | 0.3281 |
| `PaymentMethod` is `Electronic check` | 0.2559 |
| `DeviceProtection` | -0.2446 |
| `SeniorCitizen` | 0.1884 |
| `PaymentMethod` is `Bank transfer (automatic)` | -0.1281 |
| `Dependents` | -0.0870 |
| `Partner` | -0.0795 |
| `PaymentMethod` is `Credit card (automatic)` | -0.0456 |
| `MonthlyCharges` | 0.0389 |
| `TotalCharges` | -0.0003 |
| Intercept | -4.0999 |

As we can see with the regression coefficients, the two variables that impacts the most our model (Logistic regression) are contract (month to month) and contract (one year). Our model has a threshold at 0.500, it allows us to have a very good percentage of prediction (77%) and 61% of precision which means that we have a lot of true positive.As we can see the ROC curve is quite good. The AUC of the model is 0.830. What we can observe is that it seems that 61% of our data would give 100% of true positive.

We can also say that the customers that have a contract for a year have more chances to stay than customers who have a month to month contract.

# Our Recommendations

What we have observed is that people that have more chances to churn are people who have low total charges. Which means they don't have contracts with engagement.

The number of services impacts the churn. Customers that subscribed to many services are less likely to leave.

To limit the churn we need to focus on customers that have the less additional services and those who have month to month contracts. They should be targeted  with marketing campaign to encourage them to subsribe to additional services

We should propose reduction on yearly contract or/and promote additional services to increase the subscription to additional options.

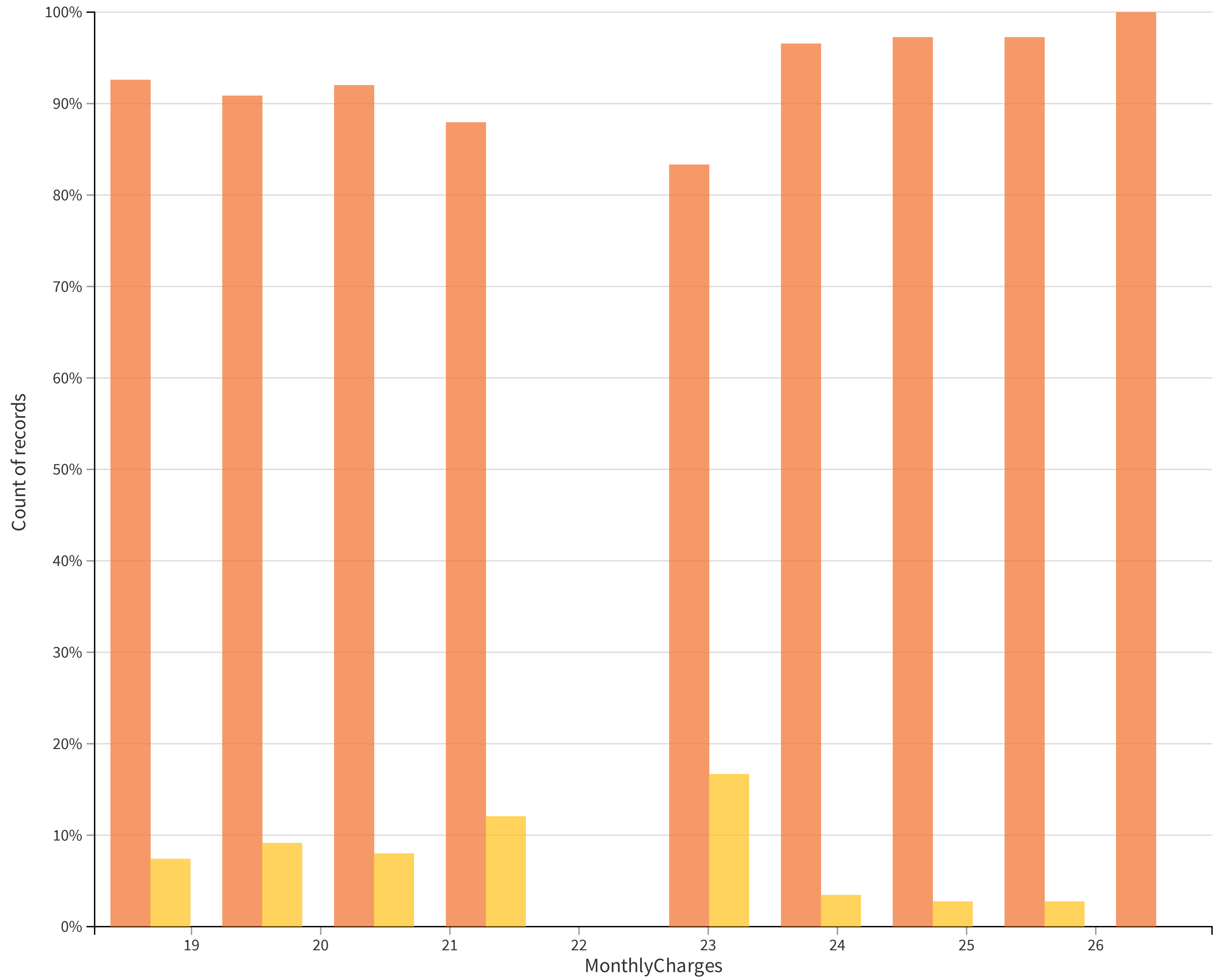
Data exploration

Phone Services customers

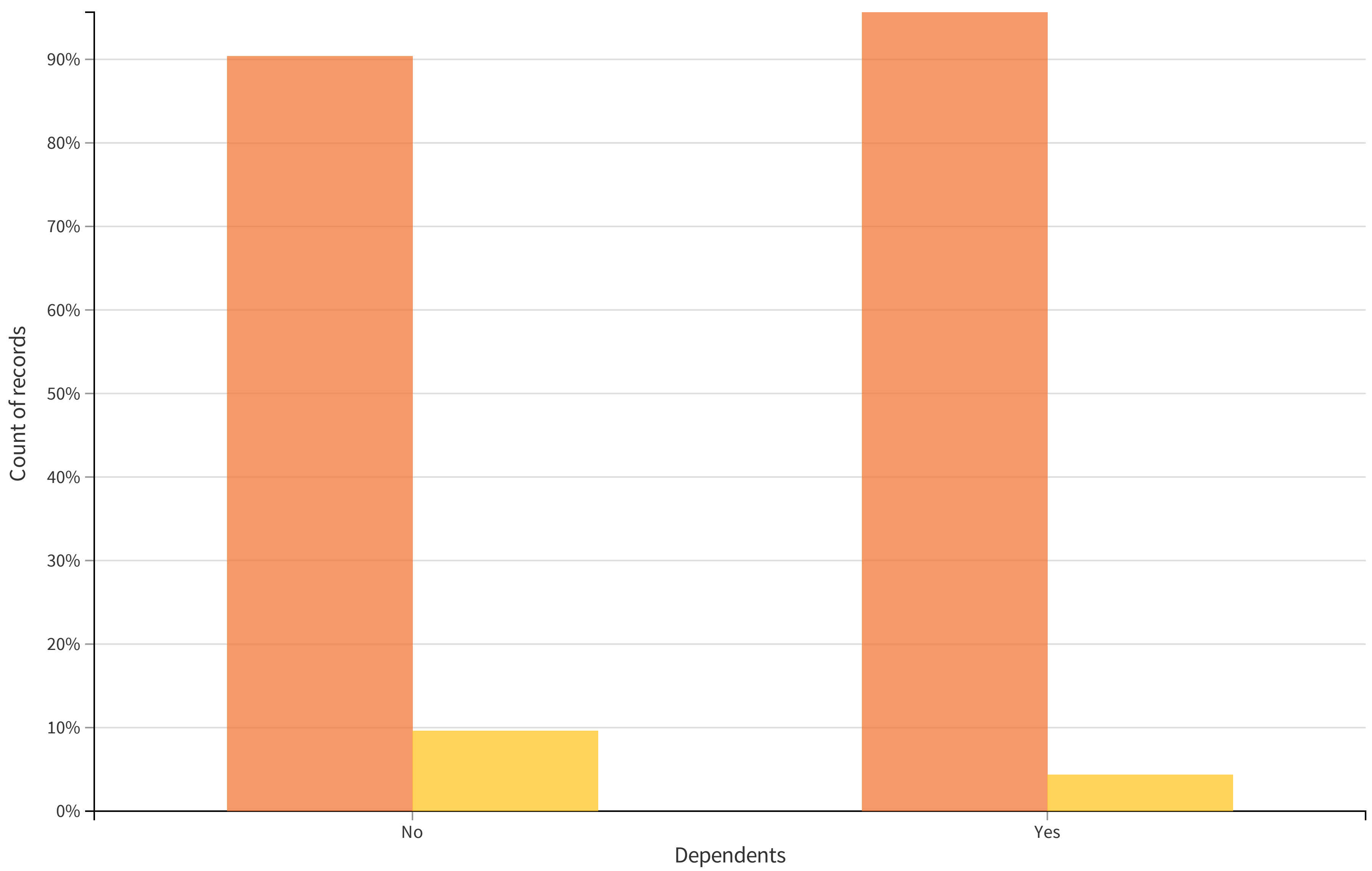
## Only phone service dataset

| customerID | gender | SeniorCitizen | Partner | Dependents | tenure | MultipleLines | Contract | Paperle |
|---|---|---|---|---|---|---|---|---|
| 7469-LKBCI | Male | 0 | No | No | 16 | No | Two year | No |
| 8191-XWSZG | Female | 0 | No | No | 52 | No | One year | No |
| 1680-VDCWW | Male | 0 | Yes | No | 12 | No | One year | No |
| 1066-JKSGK | Male | 0 | No | No | 1 | No | Month-to-month | No |
| 7310-EGVHZ | Male | 0 | No | No | 1 | No | Month-to-month | No |
| 9867-JCZSP | Female | 0 | Yes | Yes | 17 | No | One year | No |
| 3957-SQXML | Female | 0 | Yes | Yes | 34 | Yes | Two year | No |
| 3170-NMYVV | Female | 0 | Yes | Yes | 50 | No | Two year | No |
| 0731-EBJQB | Female | 0 | Yes | Yes | 52 | No | One year | Yes |
| 8028-PNXHQ | Male | 0 | Yes | Yes | 62 | Yes | Two year | Yes |
| 3887-PBQAO | Female | 0 | Yes | Yes | 45 | Yes | One year | Yes |
| 0318-ZOPWS | Female | 0 | Yes | No | 49 | No | Two year | Yes |
| 1862-QRWPE | Female | 0 | Yes | Yes | 48 | No | Two year | No |
| 2796-NNUFI | Female | 0 | Yes | Yes | 46 | No | Two year | Yes |

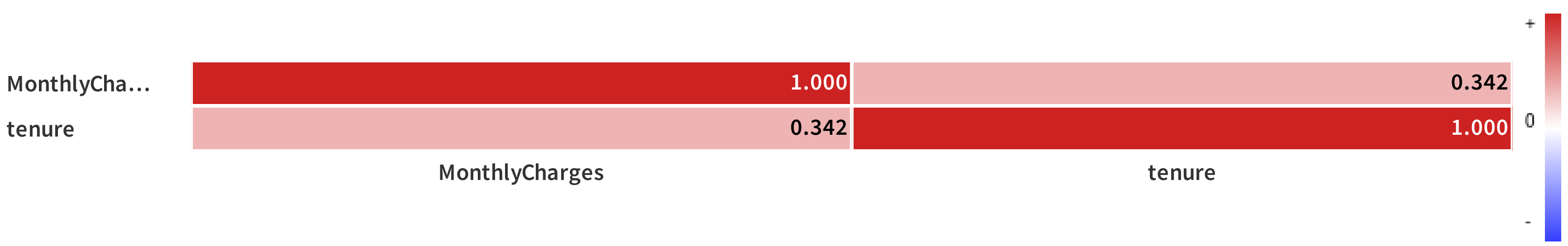
## MonthlyCharges and Churn




## Dependents and Churn




## Correlation matrix on 2 variables (Pearson)



**Observations**

- Into the database, we decide to delete all the additional services because they have to have an internet service contrat to subscribe for the options.
- As we can see, the amount of monthly charges impact the churn negatively which means that higher monthly charge is, less is the churn.
- The histogram shows that the dependents customers do not have a huge impact on the churn but we can notice that the churn is two times bigger when they are dependant.

Only phone service ML analysis

## Regression coefficients

EXPORT

Sort: | Coefficient | Filter

Display coefficients for the unscaled variables

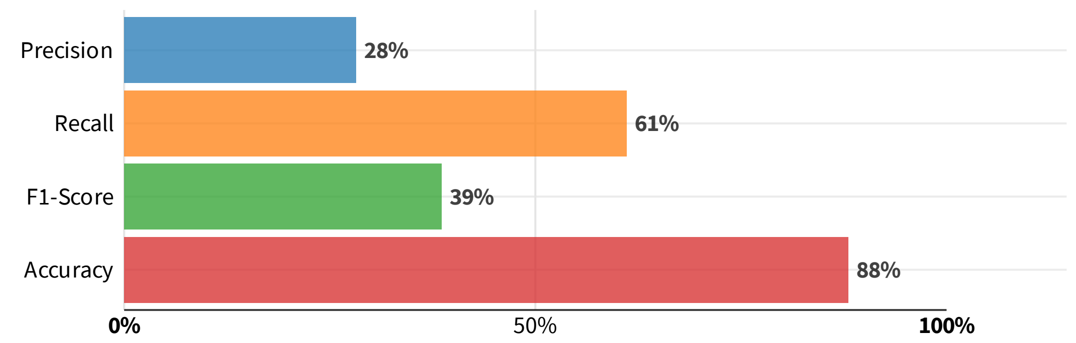
| Variable | Coefficient |
|---|---|
| Contract is Month-to-month | 0.5133 |
| Contract is Two year | -0.3838 |
| SeniorCitizen | 0.2351 |
| PaperlessBilling is No | -0.2130 |
| PaymentMethod is Credit card (automatic) | -0.2105 |
| MultipleLines is No | -0.1143 |
| Partner is No | 0.1108 |
| PaymentMethod is Mailed check | -0.0974 |
| Dependents is No | 0.0778 |
| gender is Male | -0.0661 |
| MonthlyCharges | -0.0657 |
| PaymentMethod is Bank transfer (automatic) | 0.0279 |
| tenure | -0.0235 |
| Intercept | 1.8797 |

## Confusion matrix

Threshold (cut-off) 0 1 0.675 BACK TO OPTIMAL*

Display: Record count

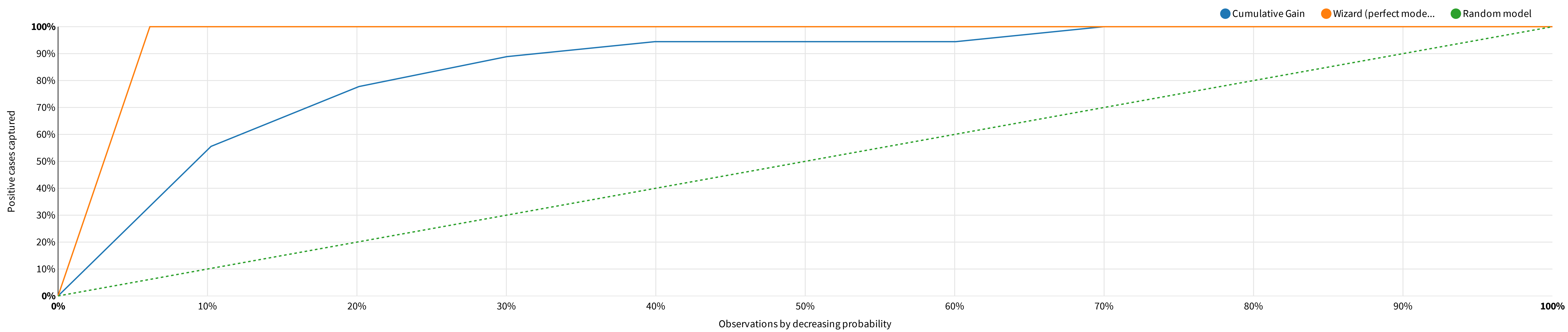
| | Predicted Yes | Predicted No | Total |
|---|---|---|---|
| Actually Yes | 11 | 7 | 18 |
| Actually No | 28 | 247 | 275 |
| Total | 39 | 254 | 293 |



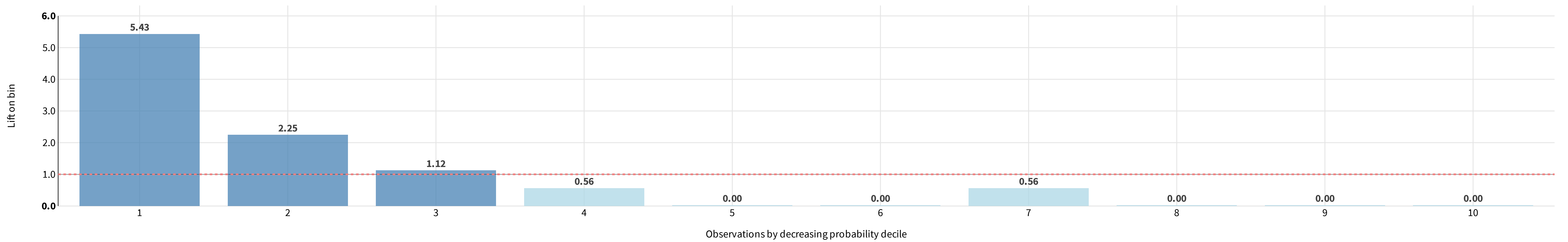

### Cost matrix

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| If model predicts Yes | and value is Yes | the gain is | 1 | × | 11 | = | 11.00 |
| | but value is No | the gain is | -0.3 | × | 28 | = | -8.40 |
| Model predicts No | and value is No | the gain is | 0 | × | 247 | = | 0.00 |
| | but value is Yes | the gain is | 0 | × | 7 | = | 0.00 |
| | **Average gain per record** | | 0.01 | × | 293 | = | 2.60 |

## Lift charts

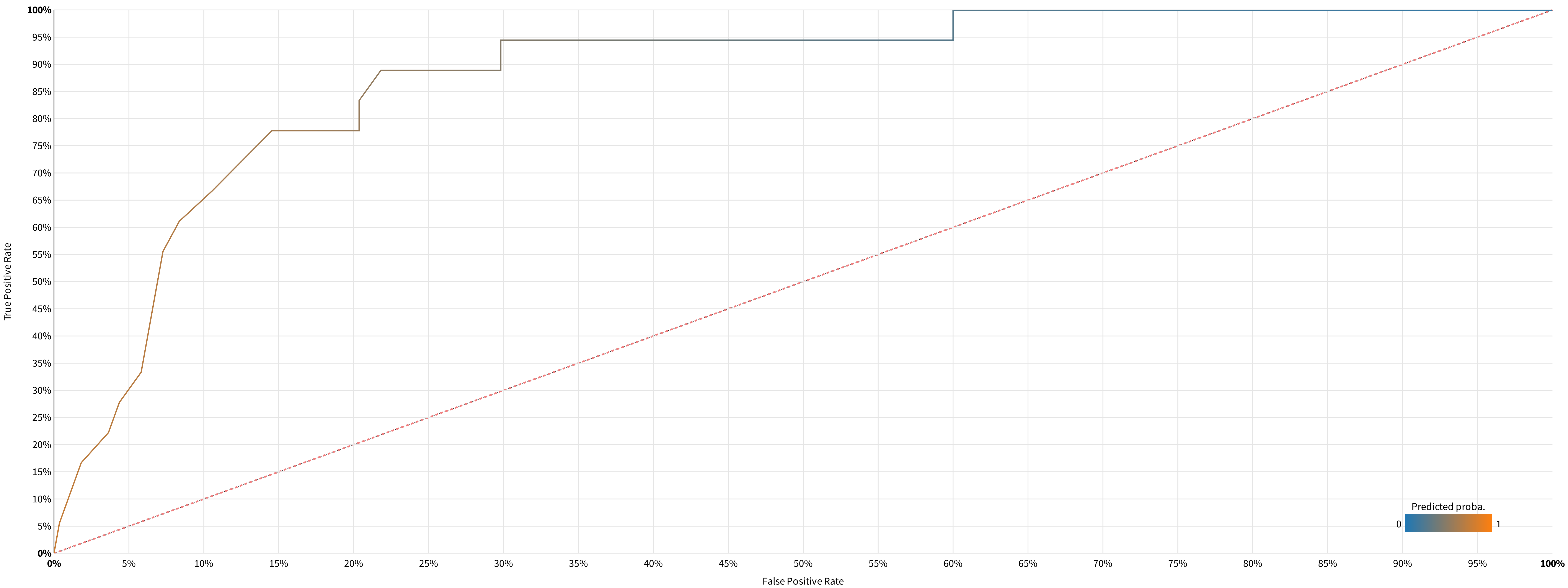


Cumulative Gain
Wizard (perfect mode...
Random model
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Positive cases captured
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Observations by decreasing probability

6.0
5.0
4.0
3.0
2.0
1.0
0.0
Lift on bin
5.43
2.25
1.12
0.56
0.00
0.00
0.56
0.00
0.00
0.00
1
2
3
4
5
6
7
8
9
10
Observations by decreasing probability decile

ROC curve



True Positive Rate
False Positive Rate
100%
95%
90%
85%
80%
75%
70%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
0%
5%
10%
15%
20%
25%
30%
35%
40%
45%
50%
55%
60%
65%
70%
75%
80%
85%
90%
95%
100%
Predicted proba.
0
1

As we can see with the regression coefficients, the two variables that impacts the most our model (Logistic regression) are contract (month to month) and contract (one year), we have the correlation with the dataset both services.

Our model has a threshold at 0.675, it allows us to have a very good percentage of recall (88%) and 61% of precision.

As we can see the ROC curve is quite good. The AUC of the model is 0.875. What we can observe is that it seems that 56% of our data would give 100% of true positive.

# Our Recommendations

We have observed that the results are very close to the "both services database" so, for a cost efficiency we decided to develop the same marketing strategy than the previous campaign.

As told before, people that have more chances to churn are people who have low total charges. Which means they don't have contracts with engagement.

The number of services impacts the churn. Customers that subscribed to many services are less likely to leave.

To limit the churn we need to focus on customers that have the less additional services and those who have month to month contracts. They should be targeted  with marketing campaign to encourage them to subsribe to additional services

We should propose reduction on yearly contract or/and promote additional services to increase the subscription to additional options.