Final report, Group 24

Group members:
Mariia Nedelkina
Juulia Sillanpää
Belinda Puutio
Nhi Nguyen

Which customers are likely to churn?

Problem:

Telecommunications company TelecomStar Oy recently launched a new product and witnessed a rise in the churn rate. The management approached us with a dataset and **the following question** to be able to get ahead and boost the user experience of potential churners:

Who's likely to churn in the near future? (i.e, the prediction problem)

Objectives:

1. To explore ML predictive models
2. To select the best-performing model for the given business case
3. To deliver the findings and code to TelecomStar, so that they can be used for similar data further on

Motivation & business value:

By answering the question and meeting the objectives, we're able to serve multiple teams of TelecomStar:

Sales team => predicted potential churners can be reached out to and offered a good deal, which helps to hit sales quotas and increase company revenue

Marketing team => better, unlikely-to-churn audience can be targeted, thus saving marketing budgets

Product team => potential churners can be researched, and the product can be improved using the findings

Dataset Overview

The dataset is sourced from Kaggle via the link

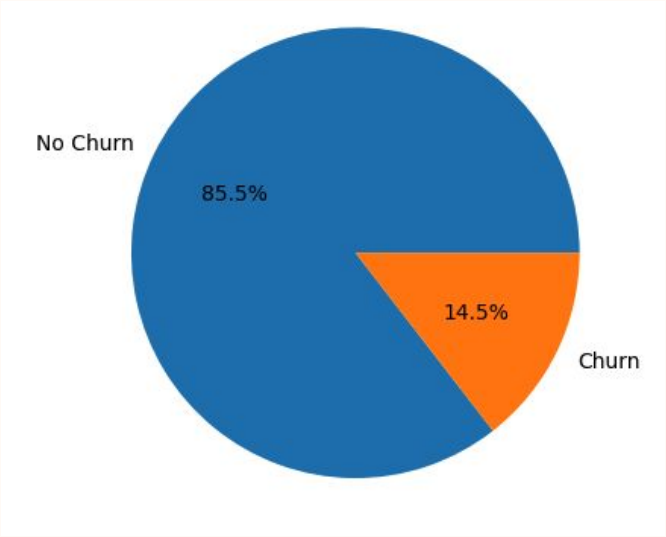
”Churn”, “ContractRenewal”, and “DataPlan” are binary and the rest are continuous variables

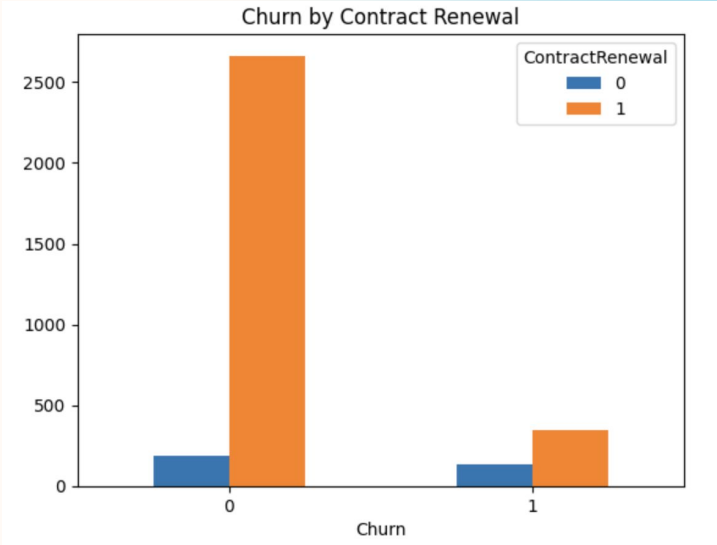
The target variable is ”Churn”

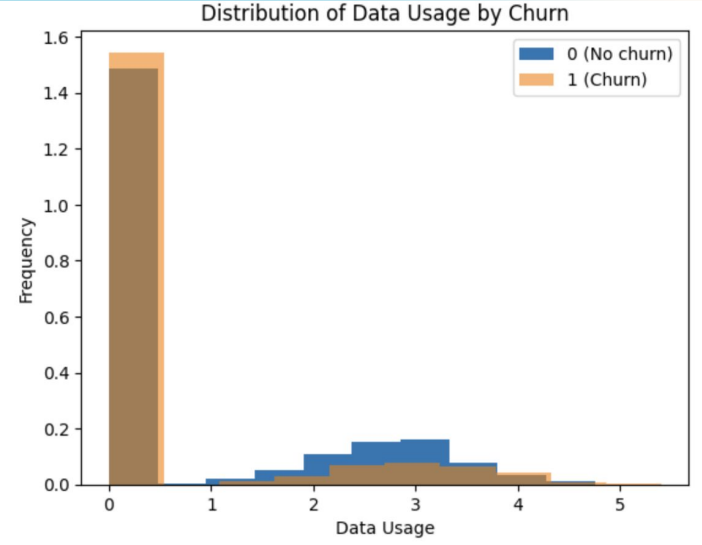
All variables are numerical so no one-hot encoding is needed

```
RangeIndex: 3333 entries, 0 to 3332
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Churn            3333 non-null   int64
 1   AccountWeeks     3333 non-null   int64
 2   ContractRenewal  3333 non-null   int64
 3   DataPlan         3333 non-null   int64
 4   DataUsage        3333 non-null   float64
 5   CustServCalls    3333 non-null   int64
 6   DayMins          3333 non-null   float64
 7   DayCalls         3333 non-null   int64
 8   MonthlyCharge    3333 non-null   float64
 9   OverageFee       3333 non-null   float64
 10  RoamMins         3333 non-null   float64
```

Descriptive analysis (1/2)







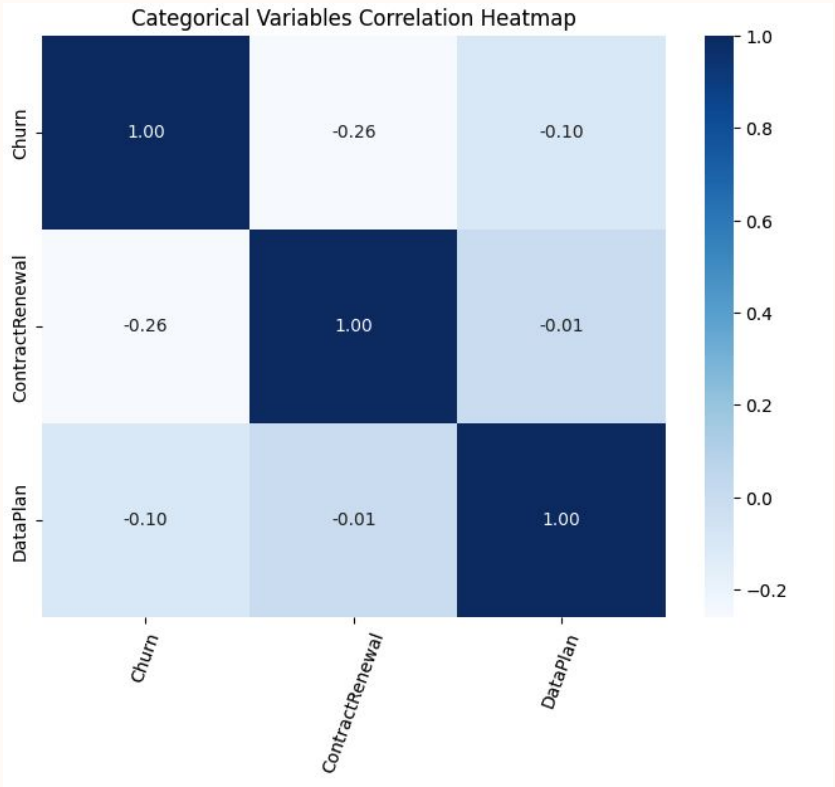
“Who’s likely to churn in the near future?” -> The goal is to predict the minority class (churners)

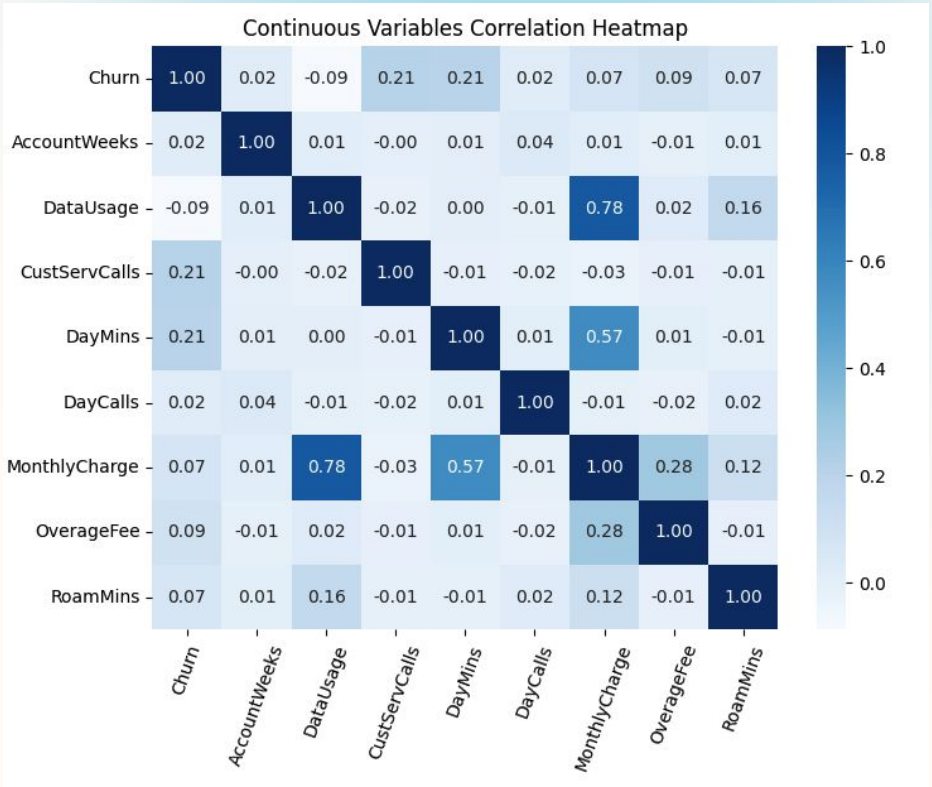
There is an imbalance in the data: 14.5% of customers churn and 85.5% don't churn

Customers who renew the contract and have more data usage are less likely to churn

Other visualizations point to a similar conclusion for users with fewer overage fees and fewer monthly charges

Descriptive analysis – correlation heatmaps (2/2)





Customer Service Calls (number of them) and DayMins (average per month) correlate the most with churn, both having a correlation coefficient of 0.21

Contract renewal in turn has the strongest inverse correlation with churn with the coefficient of –0.26

Data manipulation & selection of methods

1. StandardScaler is used to scale feature data into a standard format

2. The dataset is split into train and test sets with a test size of 30%

3. The goal is to choose the best model out of the Decision Tree, Logistic Regression, and Support Vector Machine using the original imbalanced data and rebalanced data using SMOTE and ADASYN methods. The reasoning for using the above-mentioned models:

Decision tree model
- Supports non-linearity
- Can provide an understandable explanation over the prediction

Logistic regression
- Good for classification problems
- Easy to interpret

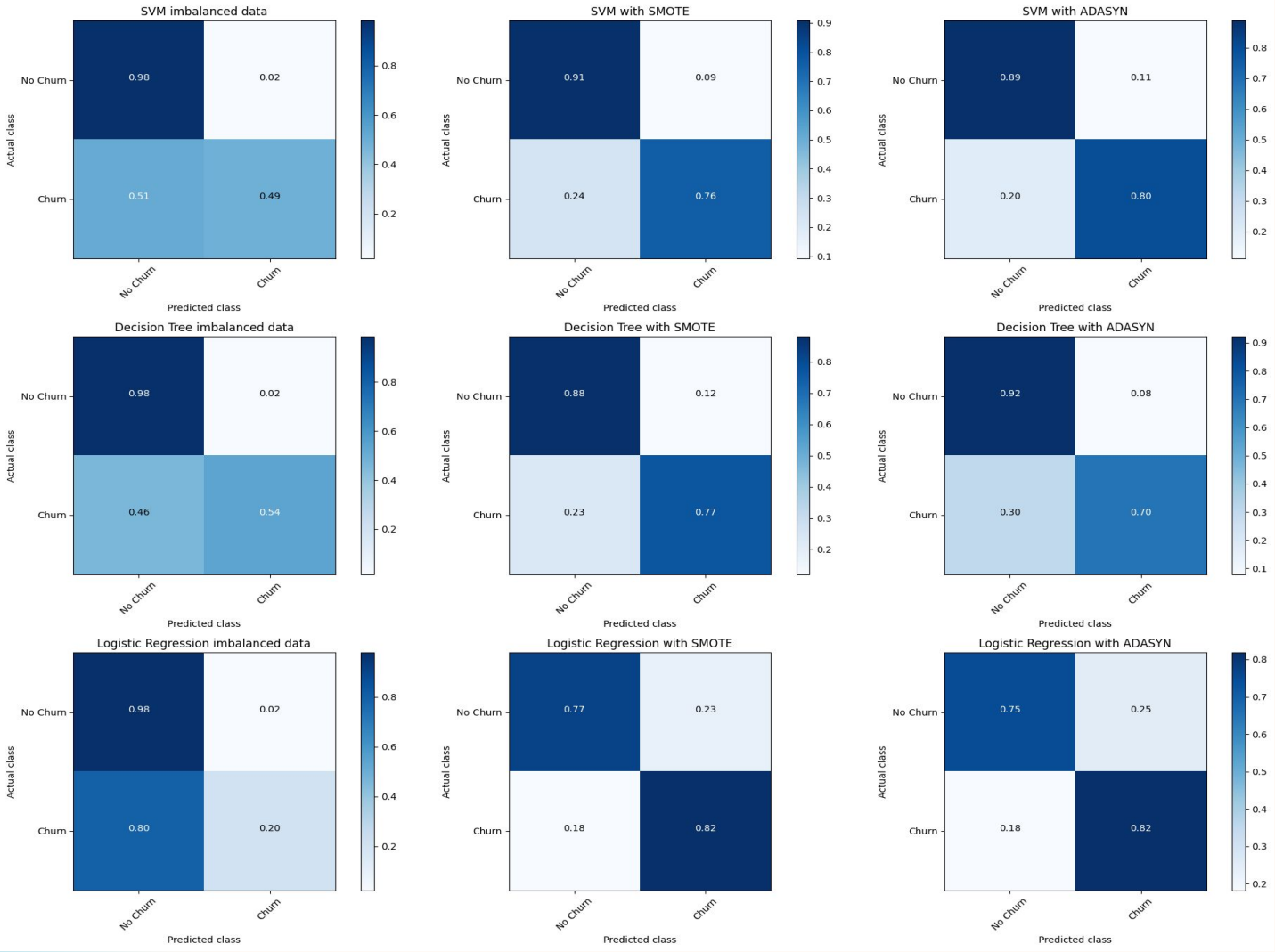
Support vector machine
- Supports both linear and non-linear solutions
- Memory efficient

Model evaluation – confusion matrices (1/3)

As can be seen from the confusion matrices, every model's **ability to predict True Positives increases significantly with balanced data.**

The best models according to True Positive rate are Logistic Regression with SMOTE, Logistic Regression with ADASYN, and SVM with ADASYN.

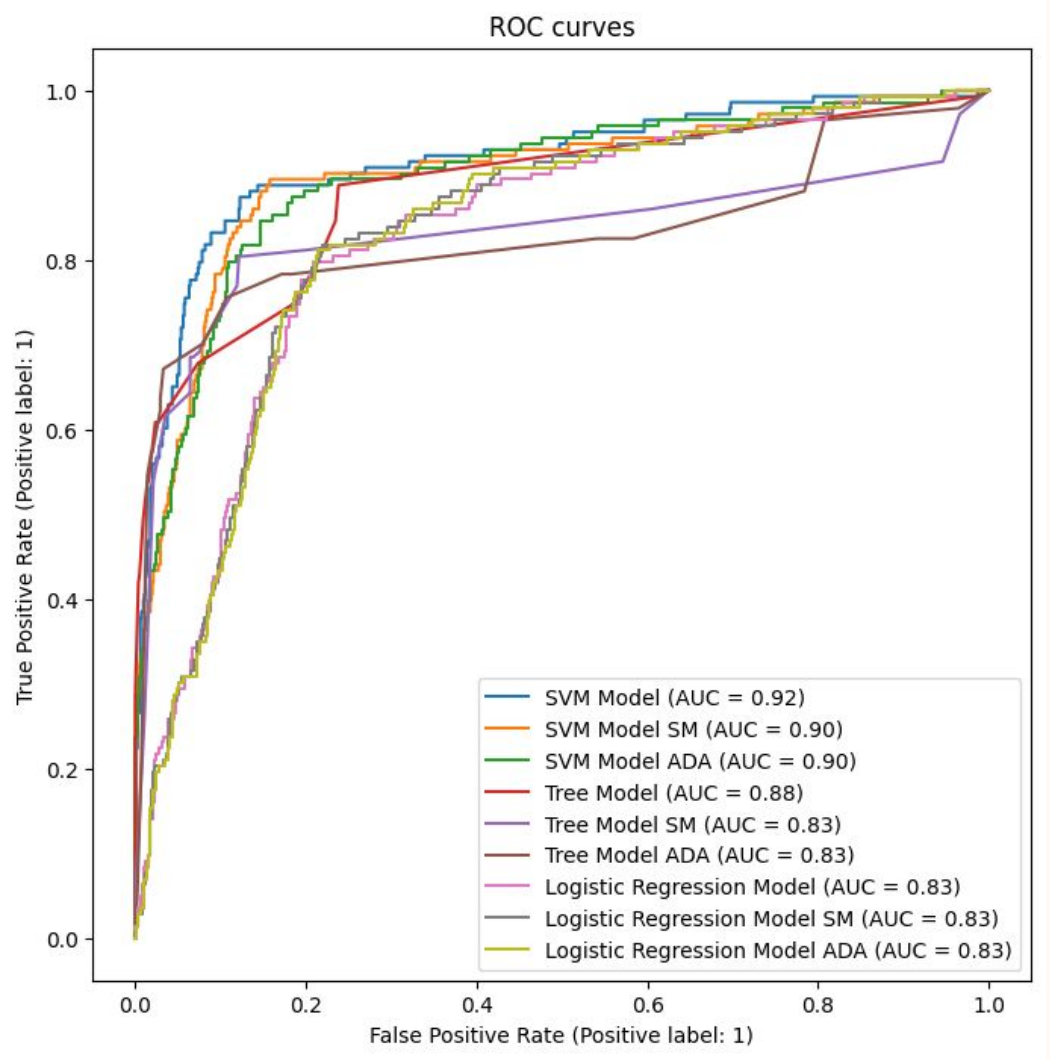
The best True Negative rates are obtained with unbalanced data.



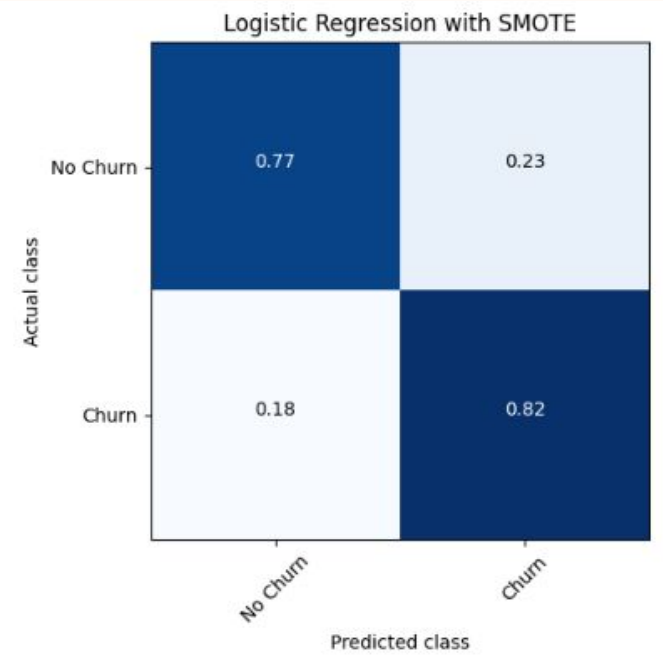
Model evaluation – ROC curves & AUC (2/3)

According to AUC, the best-performing model is SVM with unbalanced data

The good performance of models with unbalanced data is explained by the fact that they have high accuracy in predicting the non-churners i.e. True Negatives



Model evaluation – model selection (3/3)



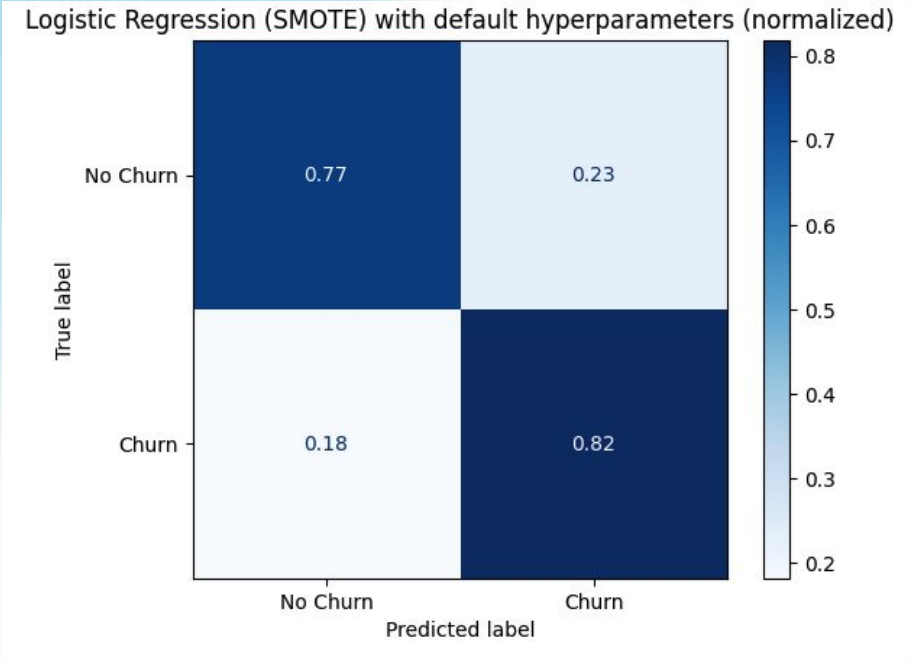
It is most important for the company to correctly predict the real churners, so that actions can be taken to prevent some of the churns. It is more costly if the actual churners go unnoticed than if the churn-preventing actions are targeted at some customers who are actually non-churners.

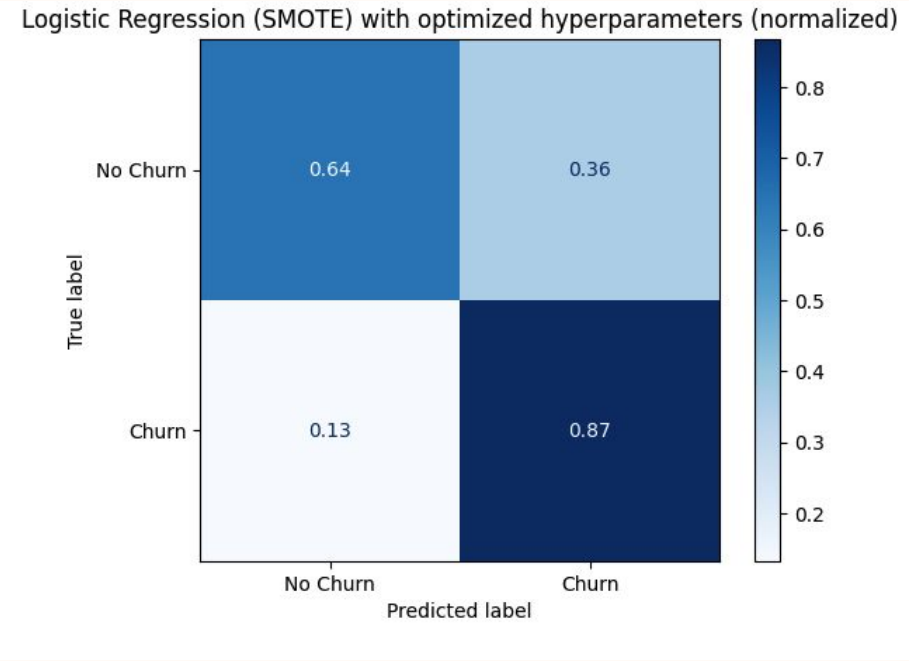
That is why we selected the final model based on the highest recall, which is telling the model's ability to correctly predict True Positives.

The selected model was **Logistic Regression, and the training data was balanced using SMOTE**.

	SVM	SVM SMOTE	SVM ADASYN	Decision Tree	Decision Tree SMOTE	Decision Tree ADASYN	Logistic Regression	Logistic Regression SMOTE	Logistic Regression ADASYN
F1	0.614035	0.658610	0.645892	0.660944	0.619718	0.645161	0.303665	0.512035	0.490566
Recall	0.489510	0.762238	0.797203	0.538462	0.769231	0.699301	0.202797	0.818182	0.818182
Accuracy	0.912000	0.887000	0.875000	0.921000	0.865000	0.890000	0.867000	0.777000	0.757000
Precision	0.823529	0.579787	0.542857	0.855556	0.518868	0.598802	0.604167	0.372611	0.350299

Hyperparameter optimization





The Logistic Regression model on resampled SMOTE dataset is optimized using Grid Search. Hyperparameter combination of {'C': 0.0001, 'max_iter': 50, 'penalty': 'l2', 'solver': 'liblinear'} yielded the highest Recall.

The Logistic Regression model after tuning has a slightly lower AUC score yet performs better at predicting customers who will churn. **Our goal is to increase the model's accuracy in finding True Positives even if the model is a bit less accurate in predicting the people who do not churn.**

Study results

Who's likely to churn in the near future? (i.e, the prediction problem)

Outcome: with the True Positive rate of 82%, the chosen model can identify potential churners

Objective 1. To explore ML predictive models

Outcome: three predictive models were examined and compared

Objective 2. To select the best-performing model for the given business case

Outcome: Logistic Regression trained on balanced data resampled by SMOTE method is the most suitable model for TelecomStar Oy

Objective 3. To deliver the findings and code to TelecomStar, so that they can be used for similar data further on

Outcome: simultaneously with this submission, the files could be delivered to the company

Extra: we're able to communicate interesting findings of descriptive analysis such as churns making more customer service calls in a month

Business recommendations

For customers at a risk of churning, run a tailored marketing campaign or offer, e.g., one month of unlimited data usage at an affordable price – to encourage customer retention

Define concrete benefits of customers staying, costs of retention efforts and build an expected value framework to target customers with the highest expected profit

Gain more insights into each feature and improve it. As customers who churn have more customer service calls, the company can do further quantitative & qualitative analysis to find insights on, e.g., what problems churning customers call the service for

For even more effective and accurate modeling, the company could also track and record other relevant demographic features that can impact the churn like salary, country/city of residence, age, and profession