Churn Prediction of 5G Boost, a Telecom Company

Artificial Intelligence Assignment 1

November 2021

Matteo Giardini

esade

Table of Contents

esade

1

Executive Summary

Executive Summary

Problem

5G Boost, a US industry leading Telco, is experiencing a **monthly customer churn rate of 14.5%:**

- At this rate, 5G Boosts' **monthly revenue** is bound to **decrease by $870,000 next month**
- According to industry standards, **customer acquisition costs (CAC) are 7x higher than retention costs**

Solution

5G Boost can use of machine learning models to **predict which customers will churn next month:**

- A dataset of **3,342 datapoints over 20 variables** with consumption data of the **past 3 months**
- **Classification models** can be used to predict exactly **which customers are likely to churn** next month
- To maximize retention, 5G Boost should **target customers with 4 specific retention initiatives**

Impact

The proposed machine learning predictive models can exactly point at **>12,000 customers who can be retained through retention initiatives.** The deployment of the model has the following effects:

- ~**24% increase in costs** (~$250,000) given by additional retention initiatives and opportunity cost of lost revenue
- ~**15% increase in revenue** (~$750,000) and ~**12% increase in next month's profits** (~$500,000)

5G Boost to avoid losing 14,500 customers next month

14.5%
monthly
churn rate

-$870,000
potential next month
revenue decrease

- According to industry standards, **acquiring a new customer is 7x higher than retaining one.**
- 5G Boost's goal should be to **maximize retention by targeting those customers who are likely to churn.**

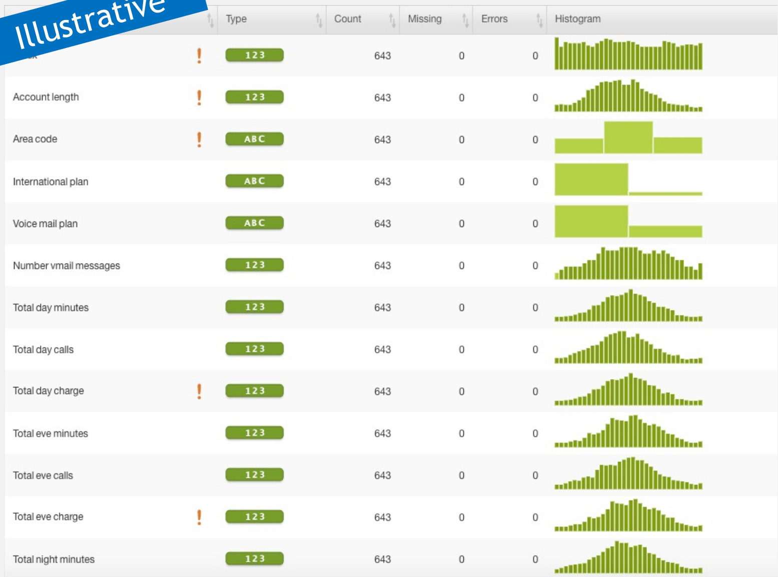
Problem | **Solution** | Impact

Supervised machine learning models to be used to predict churning customers

Detailed financial impact in following slides

Customer consumption data

Illustrative



	Type	Count	Missing	Errors	Histogram
	123	643	0	0	
Account length	123	643	0	0	
Area code	ABC	643	0	0	
International plan	ABC	643	0	0	
Voice mail plan	ABC	643	0	0	
Number vmail messages	123	643	0	0	
Total day minutes	123	643	0	0	
Total day calls	123	643	0	0	
Total day charge	123	643	0	0	
Total eve minutes	123	643	0	0	
Total eve calls	123	643	0	0	
Total eve charge	123	643	0	0	
Total night minutes	123	643	0	0	

- **3,432 datapoints** distributed over 20 features
- **3-month backward-looking consumption data**
- Useful to train **classification model** to predict customers likely to churn next month

Model output

Cost per customer

Actual \ Predicted	Stays	Churns
Stays	$0	-$10
Churns	-$60	-$10

Prediction outcome

Actual \ Predicted	Stays	Churns
Stays	85,500	0
Churns	2,072	12,428

- Economic value is assigned to each outcome
- The model is able to identify **12,428 customers who are likely to churn** next month

97.8% accuracy[1]

85.7% recall[2]

100% precision[3]

esade

Definitions in business terms: 1. Accuracy: How likely the model makes mistakes in predicting churn, 2. Recall: How well is the model effectively predicting churn overall, 3. Precision: How trustable is the model in predicting churn

Retention may be maximized through 4 main initiatives

	Initiative	Description
1	Flat charges for intensive users	• Special offers for intensive users (i.e., with high total charges) • Monthly spending not too exceed $50 • If $50 threshold is reached, 20% discount it offered for the following month
2	Special attention to customer service calls	• Intensive follow-ups and pre-emptive retention efforts for customers with more than 3 recent customer service calls
3	International plan for low-spending customers	• Begin offering a special international plan for low-spending customers • Allow these customers to make international calls occasionally for a convenient price
4	Day-caller and night-caller special packs	• Begin offering day-caller and night-caller packs with unlimited calls for a premium price respectively during the day (9am to 6m) or night (9pm to 3am)

Problem | Solution | **Impact**

Deployment of the model results in a 12% increase in profits next month

$750k
increase in revenue next month

+15%

–

$250k
increase in costs next month

+24%

=

$500k
increase in profits next month

+12%

- By deploying the newly developed machine learning model, 5G Boost will be able to **retain more than 85% of those customers who were likely to churn** and **increase monthly profits by 12**%
- Additionally, 5G Boost could leverage the insights obtained through this machine learning model to **refine its offerings** and further **increase customer retention**

2

Answers To Exercises

Definition of the Business Problem

1 State the problem in business language. What do we want to improve?

“5G Boost is currently experiencing a monthly customer churn rate of 14.5%. In order to be able to maximize its profits, the company wishes to analyse its 3-month historical consumption data to identify those customers who are likely to leave the company. 5G Boost’s goal is to maximize retention by engaging customers through personalized strategies and solidify its long-term relationships with its customer base.“

This problem can be solved with Supervised Learning

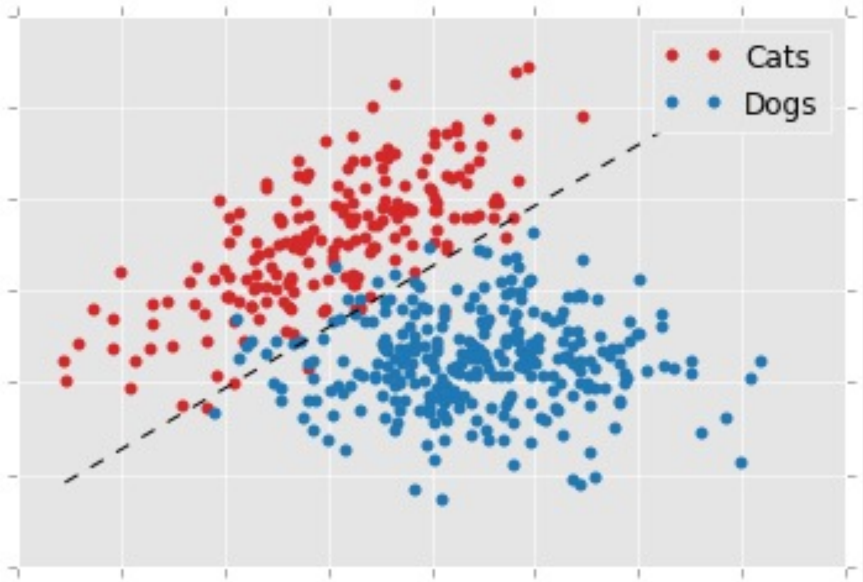
2 ML looks useful to solve this problem. Which kind of ML model do we want to train? Supervised or Unsupervised? If Supervised, is it a Classification or a Regression problem? Why?

Why Supervised Learning?

- A number of **independent variables** (or features) may be utilized as input to **predict a dependent variable** (i.e., target)
- Relationships between variables may be identified to **develop an approximate function to predict the dependent variable** (i.e., objective field) as accurately as possible
- For example, the total amount of minutes or calls (per day, evening or night) and the corresponding prices may cause customers to leave the company for reasons such as excessive prices. Thus, a **relationship may be uncovered between the different variables to predict whether customers churn or not.**

Why Classification?

As the **dependent variable (*Churn*) is categorical (True or False)**, this problem is solvable through **classification** (as shown below)



Interpretability is key

3 Is the interpretability of the ML model important in this context? Why?

Interpretability of this Machine Learning Model is important in this specific context for **three reasons:**

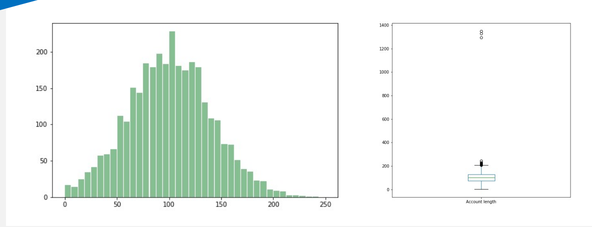
- To properly evaluate and understand the **accuracy of the model**
- To present the results obtained transparently and inform/**influence top management decision making**
- To fundamentally understand how to **improve the underlying algorithms** and finetune prediction

Data Preparation and Sanity Check

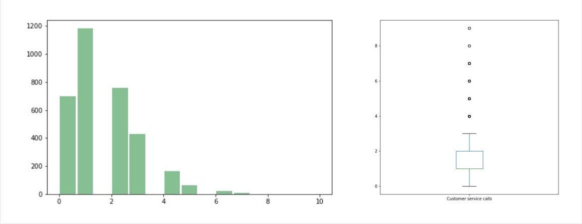
4 Sanity check: Does any variable have a strange distribution? Are there suspicious/corrupted values? Are there missing values or outliers?

Illustrative

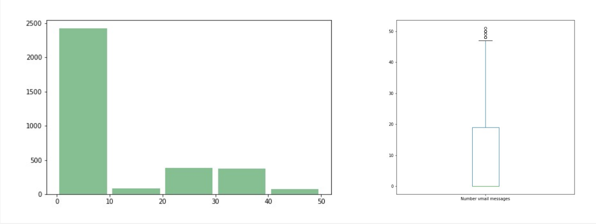
Account Length



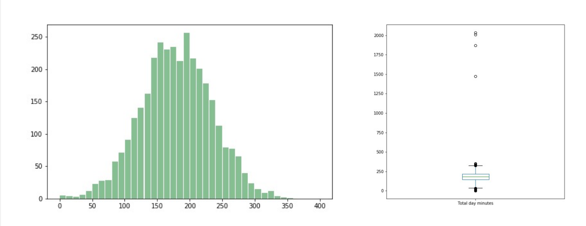
Customer Service Calls*



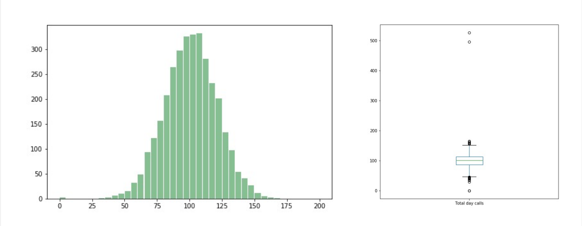
Number of Voicemail Messages*



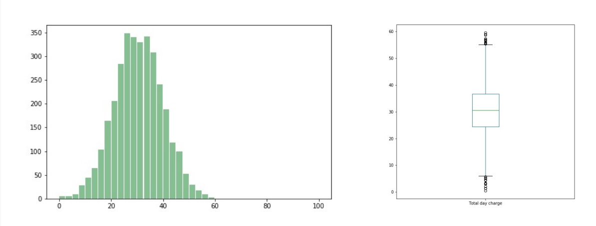
Tot Day Minutes



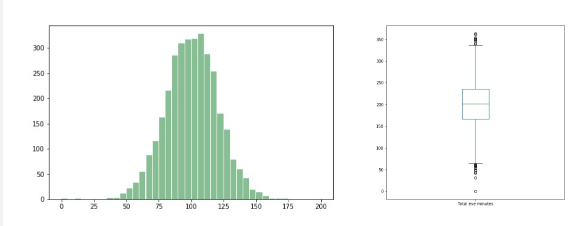
Tot Day Calls



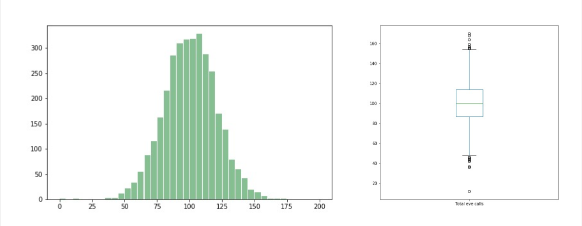
Tot Day Charge



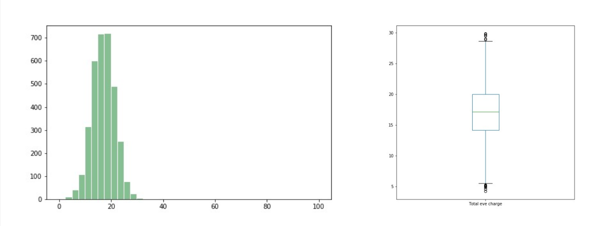
Tot Eve Minutes



Tot Eve Calls



Tot Eve Charge



Sanity Check: histograms on the left depict the distribution of the variables in the data set (not exhaustive)

Outliers: were removed from the dataset if 3 standard deviations from mean (z > 3)[1]

Missing Values: No missing values were detected in the dataset

Total count after data preparation:
3212

1. Operation performed on Account Length, Total minutes, Calls and Charge (day, evening, night and international)

* Outliers have not been removed

In the data set provided, Target Leakage does not occur

5 Is there any feature causing a Target leakage?

Definition of Target Leakage[1]	**Target leakage** occurs when a variable that is not a feature is used to predict the target. This occurs when the model is built, or trained, with information (known as the training dataset) that will not be available in unseen data

In the provided dataset, at least at first sight, **Target Leakage does not seem to occur**. All features included in the dataset seem to be information gathered before a customer could decide to leave the company:

- Categorical variables such as '*Area Code*', '*Intl Plan*', '*Voice Mail Plan*' are relevant in predicting Churn
- *Total minutes, calls and charge* for the different times of the day may hide important patterns and relationships and they should be already available prior to the customer's decision to churn
- Similar conclusions may be stated for other numerical variables such as '*Number vmail messages*' and '*Customer service calls*'

Source 1: http://downloads.alteryx.com/betawh_xnext/MachineLearning/MLTargetLeakage.htm

Define Goals | **Data Preparation** | Create Model | Interpret Model | Implement Model

‘State’ to be removed from the dataset

6 Are all the features useful to predict the target variable?

Illustrative

State	Account length	Area code	International plan	Voice mail plan	Number vmail messages	Total day minutes	Total day calls	Total day charge	Total eve minutes	Total eve calls	Total eve charge	Total night minutes	Total night calls	Total night charge	Total intl minutes	Total intl calls	Total intl charge	Customer service calls	Churn
KS	128.0	415	No	Yes	25.0	265.1	110.0	45.07	197.4	99.0	16.78	244.7	91.0	11.01	10.0	3.0	2.70	1.0	False
OH	107.0	415	No	Yes	26.0	161.6	123.0	27.47	195.5	103.0	16.62	254.4	103.0	11.45	13.7	3.0	3.70	1.0	False
NJ	137.0	415	No	No	0.0	243.4	114.0	41.38	121.2	110.0	10.30	162.6	104.0	7.32	12.2	5.0	3.29	0.0	False
OH	84.0	408	Yes	No	0.0	299.4	71.0	50.90	61.9	88.0	5.26	196.9	89.0	8.86	6.6	7.0	1.78	2.0	False
OK	75.0	415	Yes	No	0.0	166.7	113.0	28.34	148.3	122.0	12.61	186.9	121.0	8.41	10.1	3.0	2.73	3.0	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
WI	114.0	415	No	Yes	26.0	137.1	88.0	23.31	155.7	125.0	13.23	247.6	94.0	11.14	11.5	7.0	3.11	2.0	False
AL	106.0	408	No	Yes	29.0	83.6	131.0	14.21	203.9	131.0	17.33	229.5	73.0	10.33	8.1	3.0	2.19	1.0	False
VT	60.0	415	No	No	0.0	193.9	118.0	32.96	85.0	110.0	7.23	210.1	134.0	9.45	13.2	8.0	3.56	3.0	False
WV	159.0	415	No	No	0.0	169.8	114.0	28.87	197.7	105.0	16.80	193.7	82.0	8.72	11.6	4.0	3.13	1.0	False
CT	184.0	510	Yes	No	0.0	213.8	105.0	36.35	159.6	84.0	13.57	139.2	137.0	6.26	5.0	10.0	1.35	2.0	False

The variable State is **not correlated with any other variable** and is rather **unlikely to hold any strong predictive power**. For this reason, it may be initially discarded.

esade

Splitting the dataset in train and test (80%-20%)

7 Perform the train-test split. Which percentages did you choose? Why?

Train Dataset
2,569 obs.

Test Dataset
643 obs.

Train-Test Split Ratio

- The **split ratio** utilized to divide the dataset into train and test is **80%-20%**, according to best practices
- Such a ratio should guarantee that the two datasets are **suitable representations of the main dataset**
- The main goal is therefore to **optimize model training** and **maximize prediction performance**

Share of Churned Customers

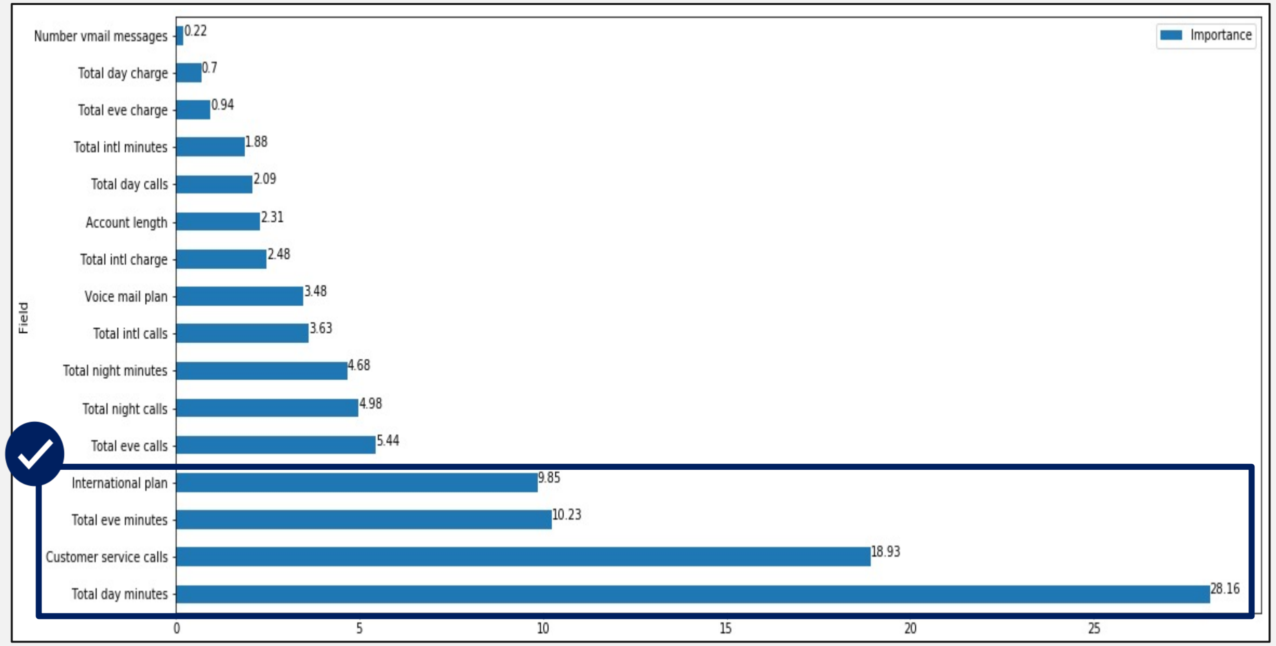
	Train	Test
False	0.854418	0.856921
True	0.145582	0.143079

- Being a small dataset with unbalanced target variables, it is fundamental that the **share of churned customers is equal** both in the train and test datasets.
- A seamlessly perfect split occurs in our case, thus it is **not necessary to perform a stratified train-test split.**

Note: Python code in Technical Appendix

Preliminary Feature Importance Analysis

8 Train a simple model and briefly analyse its metrics and the feature importance.



Feature importance above 10%

- *International plan*, *Total eve minutes*, *Customer service calls* and *Total day minutes* seem to be the most relevant features to predict Churn

Feature importance below 10%

- All those fields with importance between 0% and 5% may be discarded to optimize prediction performance

Predicted Distribution

	%	# instances
False	86.42	2,081
True	13.58	327

Trial 1

Accuracy	F-measure	Precision	Recall	Phi-coefficient
91.9%	0.7263	74.7%	70.7%	0.6789

By giving a quick look at the performance metrics, it may be observed that whilst this first model classifies instances correctly (**high accuracy**), it still **lacks precision** and **recall could also be substantially improved**.

Feature Engineering to increase predictive power

9 Feature engineering: Can new features be created by combining the original ones? If so, why do you think that the new ones can have a higher predictive power?

Feature Engineering on original dataset

In total, **three columns were added and one modified:**

1. **Total minutes, total calls and total charge** were created by adding the respective columns for for day, eve, night and intl
2. **Customer service calls** was transformed into a **categorical** variable

Why?

1. The new features might have **more predictive power** as they may add more value by giving an **overarching view** on customer's total spending, total calls and time spent on the phone.
2. By segmenting customer service calls into 2 buckets (0-2, 2-4 and 4+) the **complexity of the model may be reduced**

New Features

Customer service calls	Total charges	Total calls	Total minutes
0-2	75.56	303.0	717.2
0-2	59.24	332.0	625.2
0-2	62.29	333.0	539.4
2-4	66.80	255.0	564.8
2-4	52.09	359.0	512.0
...	...	...	...
2-4	50.79	314.0	551.9
0-2	44.06	338.0	525.1
2-4	53.20	370.0	502.2
0-2	57.52	305.0	572.8
2-4	57.53	336.0	517.6

Trial 3

Accuracy	F-measure	Precision	Recall	Phi-coefficient
95.0% ↑	0.8095 ↑	82.9% ↑	79.1% ↑	0.7812 ↑

By training another simple model **ex-post Feature Engineering**, it may be observed that **all performance metrics have improved,** hence it may be concluded that the new features have **added predictive power.**

esade

Feature selection to improve predictive performance

10 Taking into account the remaining variables, if you find that removing some more improves and simplifies your model, do feature selection.

Based on Importance and Feature Engineering

Trial 3

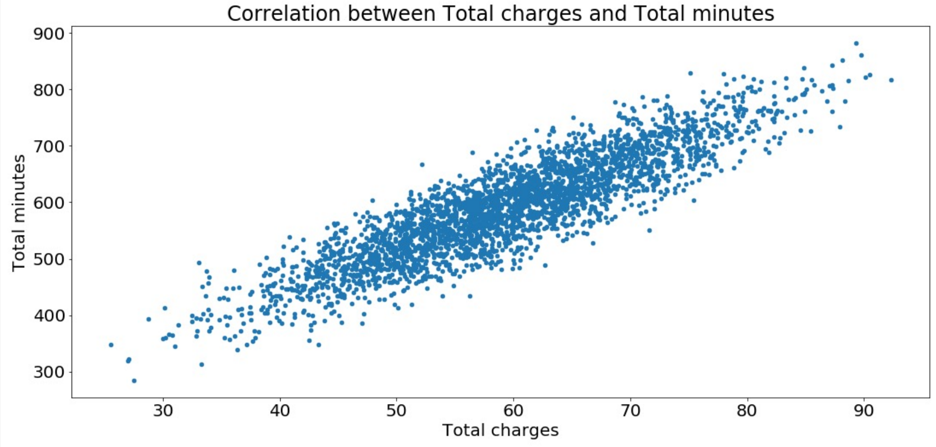
Field	Importance
Number vmail messages	0.04445
Account length	0.0174
Total night minutes	0.01153
Total intl minutes	0.00784
Total day minutes	0.00701
Total minutes	0.00699
Total calls	0.00666
Area code	0.00422

Features removed

The following features are removed:
- Account length
- Area Code
- Number of vmail messages
- Total Minutes
- Total Calls

Despite being recently added after first iteration of feature engineering

Based on Correlation



Given their **high degree of correlation with charge-related features**, all minute-related features have been removed, namely:
- Total day minutes
- Total eve minutes
- Total night minutes
- Total intl minutes

Outcome of Feature Selection

11 If applicable, did the new features or the removal of some improve the performance of your model?

	Accuracy	F-measure	Precision	Recall	Phi-coefficient
Trial 3	95.0%	0.8095	82.9%	79.1%	0.7812
Trial 4	97.2% ⬆	0.8846 ⬆	98.6% ⬆	80.2% ⬆	0.8748 ⬆

After performing the feature selection as described in the previous slide, **all performance metrics have improved.**

Accuracy	• Accuracy was already satisfactory and increased even further (by 2.2%)
Precision	• Precision was substantially improved, currently 98.6%, approaching 100%, which can be quickly reached by tweaking the T threshold in the ROC curve
Recall	• Recall was increased only slightly, perhaps further Feature Engineering is required

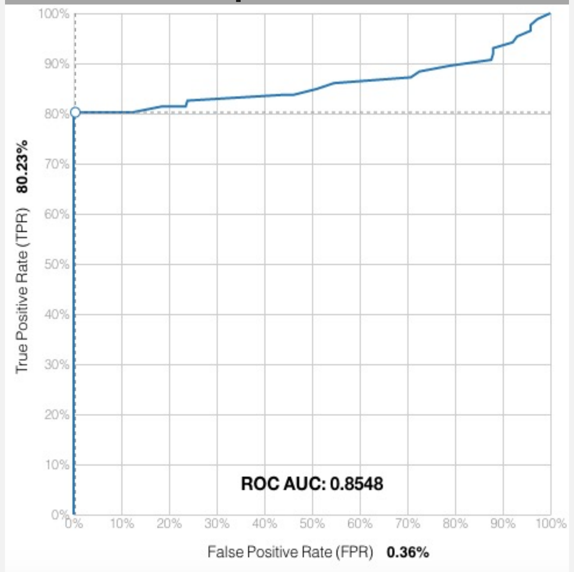
Comparison between different models (I/II)

12 Train another model with a different algorithm and compare their performance.

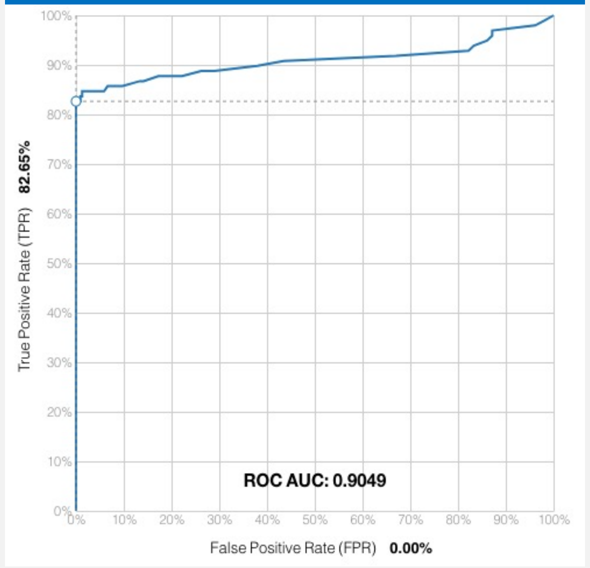
	Accuracy	F-measure	Precision	Recall	Phi-coefficient
Trial 4 Simple Model	97.2%	0.8846	98.6%	80.2%	0.8748
Trial 8 Ensemble NEW	97.4% ⬆	0.9050 ⬆	100.0% ⬆	82.7% ⬆	0.8953 ⬆

ROC curve comparison

Simple Model



Ensemble



By training an ensemble it may be noticed that **all performance metrics have been slightly improved.**

Notably:

- **Precision** reaches 100%
- **Recall** is further improved, to 82.7%

In order to further minimize False Negatives, recall may be further increased at the expense of precision.

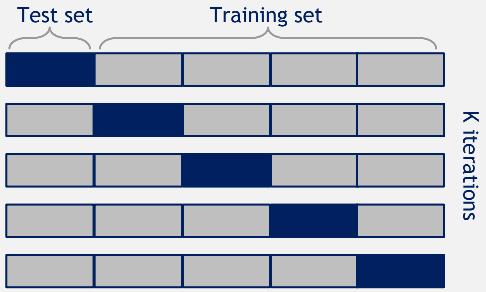
esade

5-fold cross validation in next slide

Comparison between different models (II/II)

12 Train another model with a different algorithm and compare their performance **[k-fold cross validation]**

What is k-fold cross validation?



Problem: when splitting the dataset between test and train, there is always a trade-off between the amount of data included in one dataset or the other

Solution: through k resampling iterations, cross validation allows using the entire dataset both for testing and training.

For the positive class (Churn = True), a 5-fold cross validation yields the results on the right:

- On average, the performance of these models is lower than before
- Contrarily, **average recall**, a key measure for this business problem, is significantly higher (85.42%)

The main goal of cross validation is to **evaluate the model's ability to make predictions with new data** that was not used during training.

- For this, model's performance is increased as all different scenarios included in the main dataset are taken into consideration

Avg Accuracy
95.36%
STD = 0.01242

Avg Precision
83.54%
STD = 0.03442

Avg Recall
85.42%
STD = 0.04371

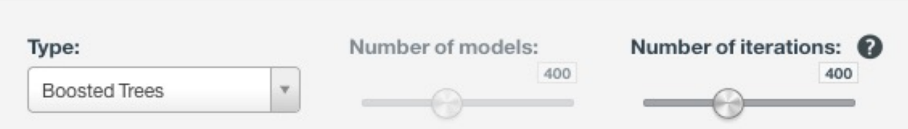
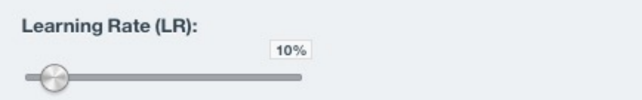
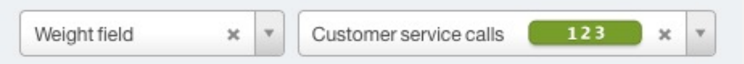
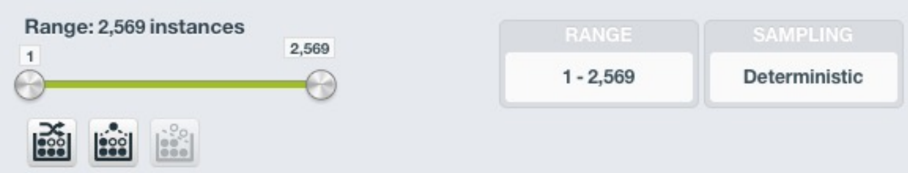
Avg F-measure
0.84
STD = 0.03093

Avg Phi
0.82
STD = 0.03774

Define Goals > Data Preparation > **Create Model** > Interpret Model > Implement Model

Finetuned Ensemble with Advanced Hyperparameters

13 Fine-tune your models and try to improve their performance.

Hyperparameters changed		Rationale
Model Type & Iterations		• Boosted Tree was selected over a simple decision tree to maintain high accuracy • 400 iterations to account for the largest # of scenarios possible
Boosting		• Early Holdout to perform the optimal number of iterations by holding out a portion of the dataset each time
Learning Rate		• Learning Rate is maintained at at 10% to avoid overfitting
Weights		• Additional weight is applied on customer service calls as it could be an early indicator of Churn
Dataset Advanced Sampling		• A deterministic sampling method is selected to use the same sampling seed, guaranteeing repeatable results

All performance metrics slights improved, notably Recall has been further increased.

Trial 9

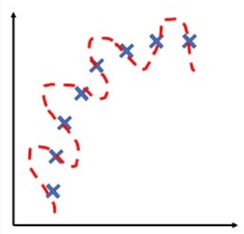
Accuracy	F-measure	Precision	Recall	Phi-coefficient
97.8% ⬆	0.9231 ⬆	100.0% ⬆	85.7% ⬆	0.9231 ⬆

esade

The model is not overfitted

14 Are your models overfitted? How did you check this? Why is an overfitted model useless?

What is overfitting and how to check

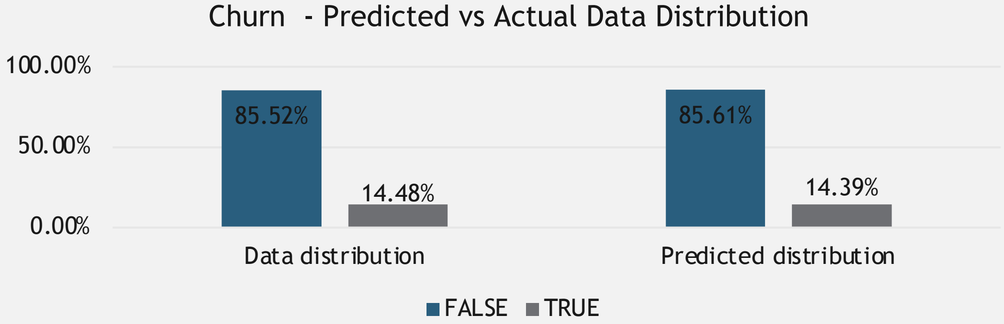


What: an overfitted model is too specific (namely not general enough) and it does not predict well for unseen data.

How: to check whether a model is overfitted it is necessary to compare the performance metrics (i.e., Precision and Recall) of the evaluation of the test and training sets.

Trial 9

	Precision	Recall
Evaluation vs Test	100.0%	85.7%
Evaluation vs Training	98.1%	86.1%



The model is not overfitted because:

- **Precision and Recall** are rather similar when the model is evaluated against the test dataset and the training dataset
- The **data distributions of the predictions and the actual dataset are the same**, thus the minority class is well represented and identified by the model

esade

Features with the highest predictive power

15 Which are the features with the highest predicting power? Why do you think that this is the case?

Trial 9

Features	Importance (%)	Possible Explanation
Total charges	23.10%	Churn is plausibly caused by excessive customer spending
Customer Service Calls	15.89%	The more a customer attempts to reach out customer service, the more likely they are to have issues which may result in churning
International Plan	13.96%	5G Boost's International plan offer might not suit customer's requirements
Total day minutes	12.04%	The more a customer uses his/her phone, the more likely they experience issues or expect special offers as a consequence of intensive usage
Total intl minutes	9.16%	Similarly to 'International plan', the company's offer might not suit customers who need to call internationally
Total night minutes	6.28%	Customers who use their phones mostly at night might expect special offers or discounts.

Confusion Matrix Interpretation

16 Interpret the Confusion Matrix of one of the models. What represents each metric (Accuracy, Recall...) and output (TP, TN...) in technical and business terms?

ACTUAL VS. PREDICTED	False	True	ACTUAL	RECALL	F	Phi
False	545	0	545	100.00%	0.99	0.91
True	14	84	98	85.71%	0.92	0.91
PREDICTED	559	84	643	92.86% AVG. RECALL	0.96 AVG. F	0.91 AVG. Phi
PRECISION	97.50%	100.00%	98.75% AVG. PRECISION	97.82% ACCURACY		

TP	Customer leaves, as predicted
TN	Customer stays, as predicted
FP	Customer stays, not as predicted
FN	Customer leaves, not as predicted

Trial 9

Accuracy	F-measure	Precision	Recall	Phi-coefficient
97.8%	0.9231	100.0%	85.7%	0.9231

	Technical Interpretation	Business Interpretation
Accuracy	# of correctly predicted instances of total prediction	How likely the model makes mistakes in predicting churn
F-measure	Balanced combination of precision and recall	Useful to compare the performance of different models
Precision	Correct predictions over <u>predicted</u> instances in positive class	How trustable is the model in predicting churn
Recall	Correct predictions over <u>actual</u> instances in positive class	How well is the model effectively predicting churn overall
Phi-coefficient	Correlation between predicted and actual values	How close are the predictions to reality

Main goal should be to maximize recall

17 Which metric would you pay more attention to if this was a real case? Why?

Recall

$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall is the most relevant metric in this case:

- **False Negatives** represent customers who are churning but that the model could not catch
- **Customer Acquisition Cost** (~$70) is **7x higher** than **Customer Retention Costs** (~$10)
- Minimizing false negatives, thus maximizing recall is essential to **limit financial damages**

F1-Score

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The **F1-Score** (or F-Measure) is:

- Useful if the goal is to **find a balance between precision and recall**
- Not essential for this problem as high precision is easily reached and, by looking at the ROC curve, **too much precision must be traded off for an increase in recall**

Economic valuation of possible model outputs

18 Based on the business interpretation of the matrix, assign a value to each possible output using the data about 5GBoost and your business reasoning. Why did you choose these values?

Cost per customer based on model outcome

Actual \ Predicted	Stays	Churns
Stays	$0	-$10
Churns	-$60	-$10

Rationale

TP	As indicated, retention efforts are 1/7 of CAC ($70)
TN	Business as usual, the model prediction is correct
FP	As indicated, retention efforts are 1/7 of CAC ($70)
FN	One churning customer means $60 loss in revenue

Deployment of the model increases profit by 12%

19 Based on the metrics of the models and the values attributed to each output, choose the model that maximizes profit. How much do you estimate that the profit of 5GBoost would change the month after the model is deployed?

Trial 9

Actual \ Predicted	Stays	Churns
Stays	85,500	0
Churns	2,072	12,428

Customers are allocated on this matrix based on test confusion matrix

Additional revenue	Thanks to the predictions, **12,428 customers are retained**, generating **14.5% increase in revenue.**
Cost of False Negatives	**$60 cost for each false negative** predicted, generating an additional **12.2% increase in costs.**
Retention Cost of True Positives	**$10 of retention costs** will be spent for each true positives to avoid churning, generating an **12.2% increase in costs.**

Current scenario without prediction model

Revenue[1]		$ 5,130,000
Costs		
Customer Acq. Cost[2]	($1,015,000)	
Profit		$4,115,000

Future scenario with prediction model

Revenue[3]		$ 5,875,680
Costs		
Customer Acq. Cost[2]	($1,015,000)	
Cost of FN[4]	($124,320)	
Cost of TP[5]	($124,280)	
Profit		$4,612,000

12% increase in profits through model deployment

Assumption: *all TP will be retained*

esade

Calculations: (1) 85,500 customers x $60, (2) 14,500 customers x $70, (3) (85,500 + 12,428 customers) x $60, (4) 2,072 customers x $60, (5) 12,426 customers x $10

Business initiatives to increase retention

20 Imagine that the model is deployed. Which business activities would you implement based on the outcomes that you receive from the model? Briefly explain some initiatives.

	Initiative	Description
1	Flat charges for intensive users	• Special offers for intensive users (i.e., with high total charges) • Monthly spending not too exceed \$50 • If \$50 threshold is reached, 20% discount it offered for the following month
2	Special attention to customer service calls	• Intensive follow-ups and pre-emptive retention efforts for customers with more than 3 recent customer service calls
3	International plan for low-spending customers	• Begin offering a special international plan for low-spending customers • Allow these customers to make international calls occasionally for a convenient price
4	Day-caller and night-caller special packs	• Begin offering day-caller and night-caller packs with unlimited calls for a premium price respectively during the day (9am to 6m) or night (9pm to 3am)

3

Appendix with Technical Procedures

Appendix: Data Preparation and Sanity Check (I/III)

4 Sanity check: Does any variable have a strange distribution? Are there suspicious/corrupted values? Are there missing values or outliers?

Code snippet example to build histograms

```
#Account Length
data['Account length'].hist(bins=40, grid=False, figsize=(8,5), range = [0,250], color='#86bf91', zorder=2, rwidth=0.9)
```

Column to be depicted is the only change from graph to graph

Code snippet example to build boxplots for entire data frame (see next slides)

```
#Box plot to evaluate which variables have to be investigated to remove outliers
data.boxplot(figsize=(28,15))
```

Code snippet example to build boxplots for each column

```
#Boxplot number of vmail messages
data.boxplot(column =['Number vmail messages'], grid = False, figsize=(8,8))
```

Column to be depicted is the only change from graph to graph

Code snippet of function to remove outliers and its application

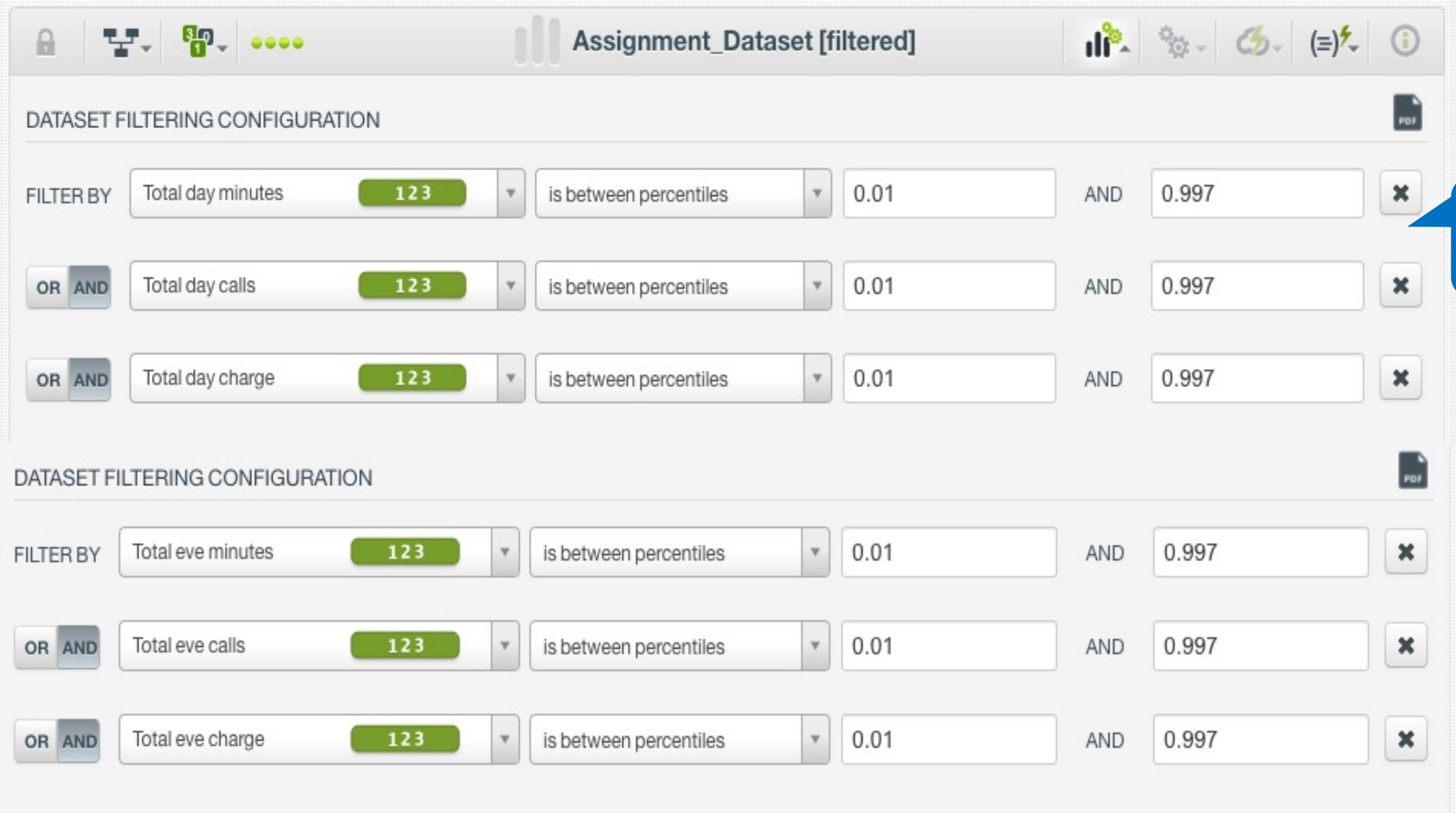
```
#If the z-score of a variable is more than 3 (meaning it is more than 3 STDs from the mean) it is considered an outlier
#define function to remove outliers given a certain threshold
def remove_outliers(df, column):
    z = np.abs(stats.zscore(df[column]))
    df = df[(z<3)]
    return df
data = remove_outliers(data, 'Total day calls')
```

esade

Appendix: Data Preparation and Sanity Check (II/III)

4 Sanity check: Does any variable have a strange distribution? Are there suspicious/corrupted values? Are there missing values or outliers?

Removal of outliers on BigML



Only observations included within the 1st and 97th percentile are included, namely those with z<3

esade

Appendix: Data Preparation and Sanity Check (III/III)

4 Sanity check: Does any variable have a strange distribution? Are there suspicious/corrupted values? Are there missing values or outliers?

Full dataset, data count, missing values and histograms with distributions on BigML

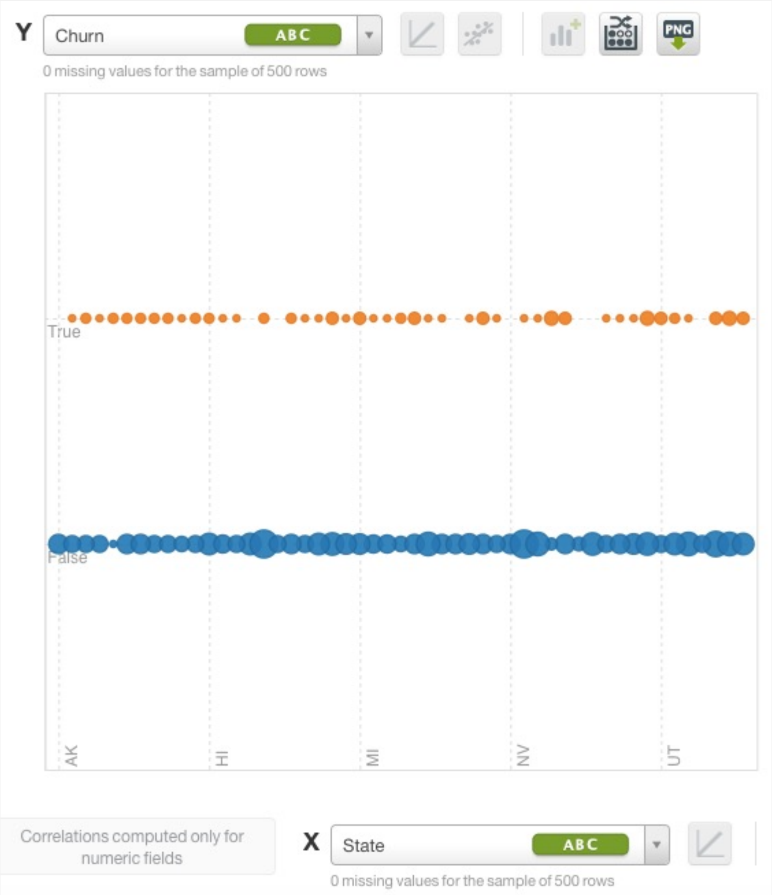
Name		Type	Count	Missing	Errors	Histogram
Index	!	123	643	0	0	
Account length	!	123	643	0	0	
Area code	!	ABC	643	0	0	
International plan		ABC	643	0	0	
Voice mail plan		ABC	643	0	0	
Number vmail messages		123	643	0	0	
Total day minutes		123	643	0	0	
Total day calls		123	643	0	0	
Total day charge	!	123	643	0	0	
Total eve minutes		123	643	0	0	
Total eve calls		123	643	0	0	
Total eve charge	!	123	643	0	0	
Total night minutes		123	643	0	0	
Total night calls		123	643	0	0	
Total night charge	!	123	643	0	0	
Total intl minutes		123	643	0	0	
Total intl calls		123	643	0	0	
Total intl charge	!	123	643	0	0	
Customer service calls		123	643	0	0	
Churn		ABC	643	0	0	
Total charges		123	643	0	0	
Total calls	!	123	643	0	0	
Total minutes	!	123	643	0	0	

esade

Appendix: Removal of State from dataset

6 Are all the features useful to predict the target variable?

Scatter plot between State and Churn



Removal of variable State

```
#Remove state
data.drop(['State'], axis=1, inplace = True)
```

Appendix: Train-Test Split

7 Perform the train-test split. Which percentages did you choose? Why?

Splitting dataset into train and test

Output of train-test split

Appendix: First Model Feature Importance

8 Train a simple model and briefly analyse its metrics and the feature importance.

Confusion Matrix and Performance Metric of 1st model (Trial 1) - Decision Tree

Trial 1

Positive class: True

TP FN FP TN

ACTUAL VS. PREDICTED	False	True	ACTUAL	RECALL	F	Phi
False	488	22	510	95.69%	0.95	0.68
True	27	65	92	70.65%	0.73	0.68
PREDICTED	515	87	602	83.17% AVG. RECALL	0.84 AVG. F	0.68 AVG. Phi
PRECISION	94.76%	74.71%	84.73% AVG. PRECISION	91.86% ACCURACY		

91.9% Accuracy	0.7263 F-measure

74.7% Precision	70.7% Recall	0.6789 Phi coefficient

Feature Importance (Trial 1)

```
df['Importance'] = (df['Importance']*100).round(decimals = 2)
df
```

	Field	Importance
0	Total day minutes	28.16
1	Customer service calls	18.93
2	Total eve minutes	10.23
3	International plan	9.85
4	Total eve calls	5.44
5	Total night calls	4.98
6	Total night minutes	4.68
7	Total intl calls	3.63
8	Voice mail plan	3.48
9	Total intl charge	2.48
10	Account length	2.31
11	Total day calls	2.09
12	Total intl minutes	1.88
13	Total eve charge	0.94
14	Total day charge	0.70
15	Number vmail messages	0.22

Bar chart with feature importance

```
bar_chart = df.plot.barh(x='Field', y='Importance', rot=0, figsize = [18,8])

for i, v in enumerate(df['Importance']):
    bar_chart.text(v, i, str(v))

plt.savefig('Simple_Model_Field_Importance.png')
plt.show()
```

esade

Appendix: Feature Engineering

9 Feature engineering: Can new features be created by combining the original ones? If so, why do you think that the new ones can have a higher predictive power?

Code snippets of column aggregation - *Total charges, total calls, total minutes*

```python
#total charges = day + eve + night + intl
total_charges = data['Total day charge'] + data['Total eve charge'] + data['Total night charge'] + data['Total intl charge']
data['Total charges'] = total_charges
#total calls = day + eve + night + intl
total_calls = data['Total day calls'] + data['Total eve calls'] + data['Total night calls'] + data['Total intl calls']
data['Total calls'] = total_calls
#total minutes = day + eve + night + intl
total_minutes = data['Total day minutes'] + data['Total eve minutes'] + data['Total night minutes'] + data['Total intl minutes']
data['Total minutes'] = total_minutes
```

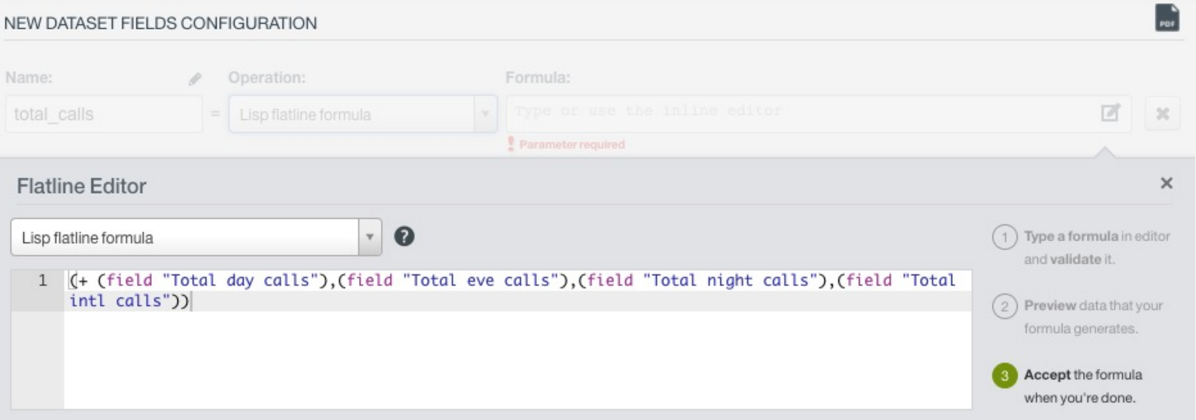
Snippet of code to turn Customer Service Calls into categories

3 buckets:
0-2, 2-4, 4+

```python
#turn Customer Service calls into categories
try:
    for i, element in enumerate(data['Customer service calls']):
        category_1 = '0-2'
        category_2 = '2-4'
        category_3 = '4+'

        if element >= 0 and element < 2:
            data['Customer service calls'].iloc[i] = category_1
        elif element >= 2 and element < 4:
            data['Customer service calls'].iloc[i] = category_2
        elif element >= 4:
            data['Customer service calls'].iloc[i] = category_3
except:
    None
```

Create Total columns on BigML



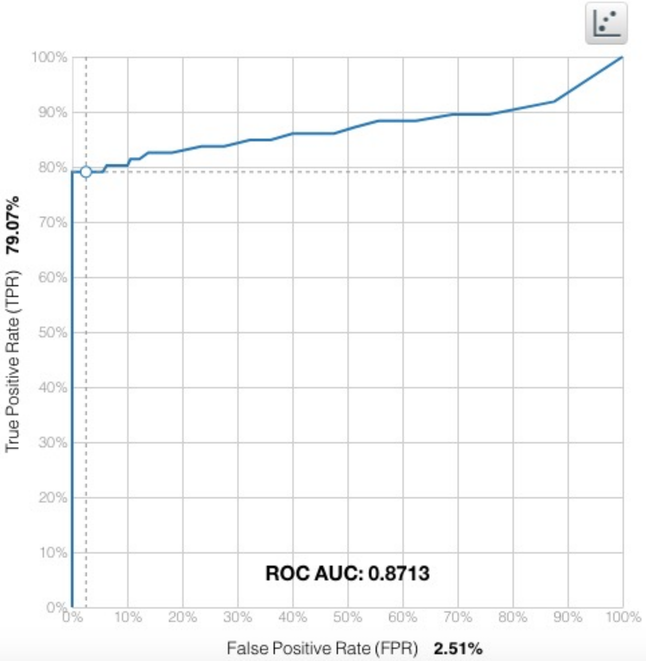
Appendix: Results of model post Feature Engineering

9 Feature engineering: Can new features be created by combining the original ones? If so, why do you think that the new ones can have a higher predictive power?

Confusion Matrix and Performance Metrics of 3rd model (Trial 3) - Post Feature Eng.

Trial 3

ACTUAL VS. PREDICTED	True	False	ACTUAL	RECALL	F	Phi
True	68	18	86	79.07%	0.81	0.78
False	14	543	557	97.49%	0.97	0.78
PREDICTED	82	561	643	88.28% AVG. RECALL	0.89 AVG. F	0.78 AVG. Phi
PRECISION	82.93%	96.79%	89.86% AVG. PRECISION	95.02% ACCURACY		



95.0% Accuracy

0.8095 F-measure

82.9% Precision

79.1% Recall

0.7812 Phi coefficient

2.5% FPR

12.8% % positive instances

620.0% Lift

79.1% K-S statistic

0.3674 Kendall's Tau

0.4387 Spearman's Rho

esade

Appendix: Feature Section based on Importance

10 Taking into account the remaining variables, if you find that removing some more improves and simplifies your model, do feature selection.

Feature Importance of Trial 3

```
Data distribution:
    False: 85.21% (2189 instances)
    True: 14.79% (380 instances)


Predicted distribution:
    False: 85.64% (2200 instances)
    True: 14.36% (369 instances)


Field importance:
    1. Total charges: 31.82%
    2. Customer service calls: 16.57%
    3. International plan: 11.27%
    4. Total intl calls: 8.72%
    5. Total intl minutes: 8.43%
    6. Total night calls: 4.59%
    7. Voice mail plan: 3.86%
    8. Total eve minutes: 3.47%
    9. Total day minutes: 2.96%
    10. Total eve calls: 2.74%
    11. Total day calls: 2.45%
    12. Total intl charge: 1.44%
    13. Total night minutes: 1.43%
    14. Total night charge: 0.26%
```

Code snippet of scatter plot between Total charges and Total minutes

```
data.plot.scatter(x='Total charges', y='Total minutes', title= "Correlation between Total charges and Total minutes", figsize = [18,8]);
plot.rcParams.update({'font.size': 10})
```

Features excluded from the model

Total day charge	!	123
Total day charge	!	123
Total eve charge	!	123
Total night charge	!	123
Total intl charge	!	123
Total calls	!	123
Total minutes	!	123

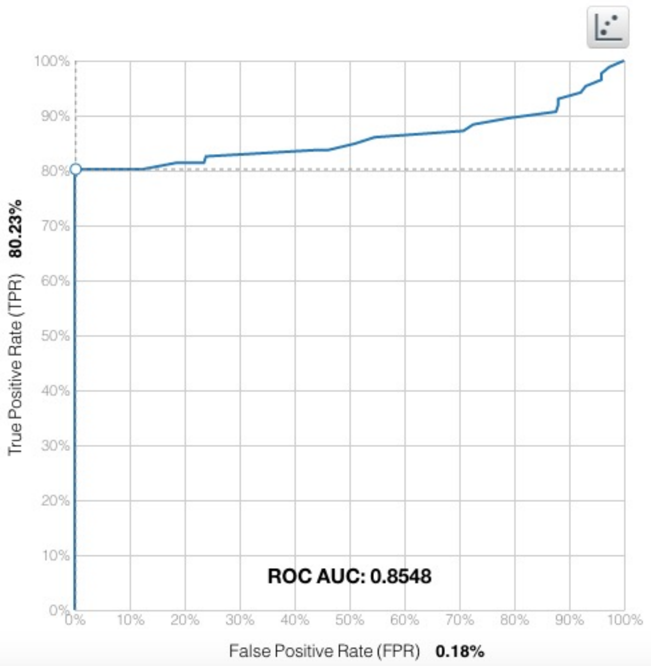
Appendix: Model post Feature Selection

11 If applicable, did the new features or the removal of some improve the performance of your model?

Confusion Matrix and Performance Metric of 4th model (Trial 4) – Decision Tree

Trial 4

ACTUAL VS. PREDICTED	True	False	ACTUAL	RECALL	F	Phi
True	69	17	86	80.23%	0.88	0.87
False	1	556	557	99.82%	0.98	0.87
PREDICTED	70	573	643	90.03% AVG. RECALL	0.93 AVG. F	0.87 AVG. Phi
PRECISION	98.57%	97.03%	97.80% AVG. PRECISION	97.20% ACCURACY		



97.2% Accuracy	0.8846 F-measure

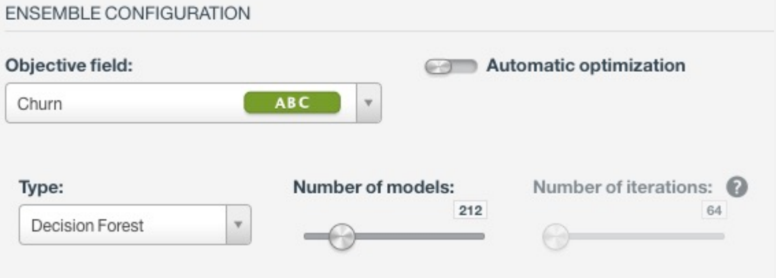
98.6% Precision	80.2% Recall	0.8748 Phi coefficient
0.2% FPR	10.9% % positive instances	737.0% Lift
80.1% K-S statistic	0.3419 Kendall's Tau	0.4183 Spearman's Rho

esade

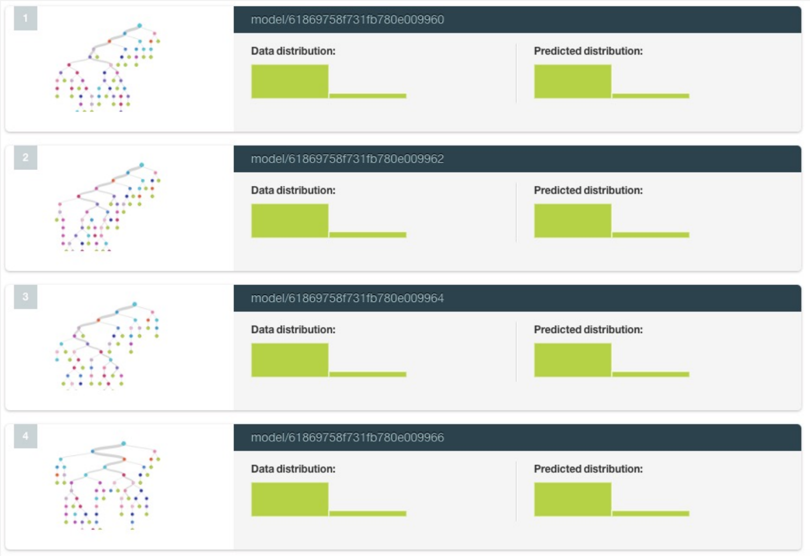
Appendix: comparison between 1st model and ensemble

12 Train another model with a different algorithm and compare their performance.

Ensemble Configuration



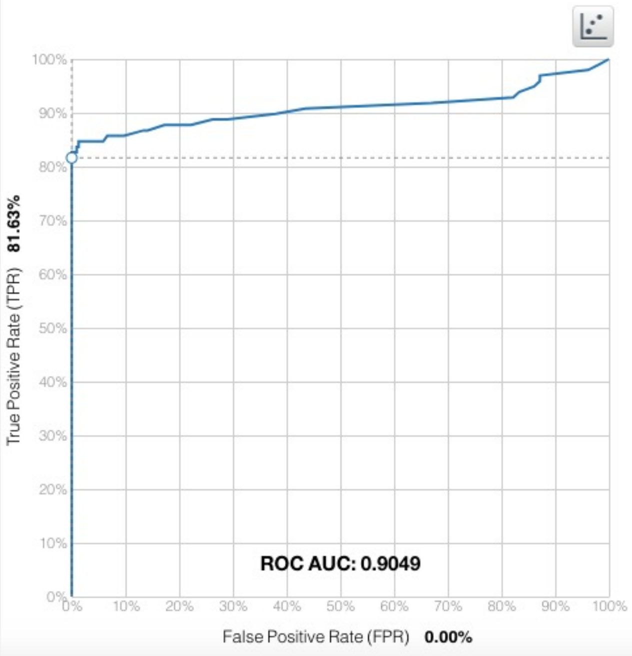
Ensemble model list



Confusion Matrix and Performance Metric of 8th model (Trial 8) - Decision Tree

Trial 8

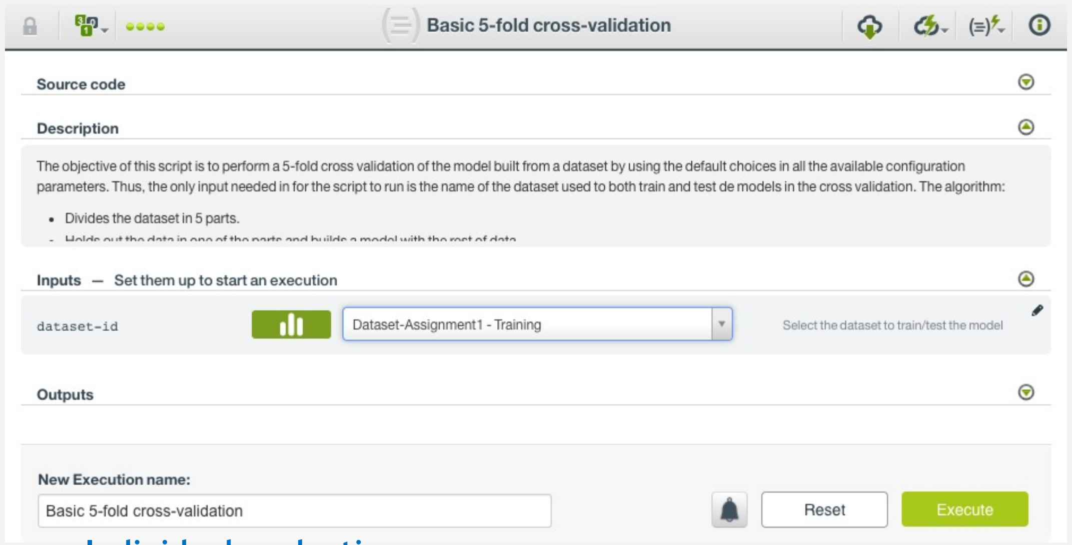
ACTUAL VS. PREDICTED	True	False	ACTUAL	RECALL	F	Phi
True	80	18	98	81.63%	0.90	0.89
False	0	545	545	100.00%	0.98	0.89
PREDICTED	80	563	643	90.82% AVG. RECALL	0.94 AVG. F	0.89 AVG. Phi
PRECISION	100.00%	96.80%	98.40% AVG. PRECISION	97.20% ACCURACY		



Appendix: 5-fold cross validation

12 Train another model with a different algorithm and compare their performance **[k-fold cross validation]**

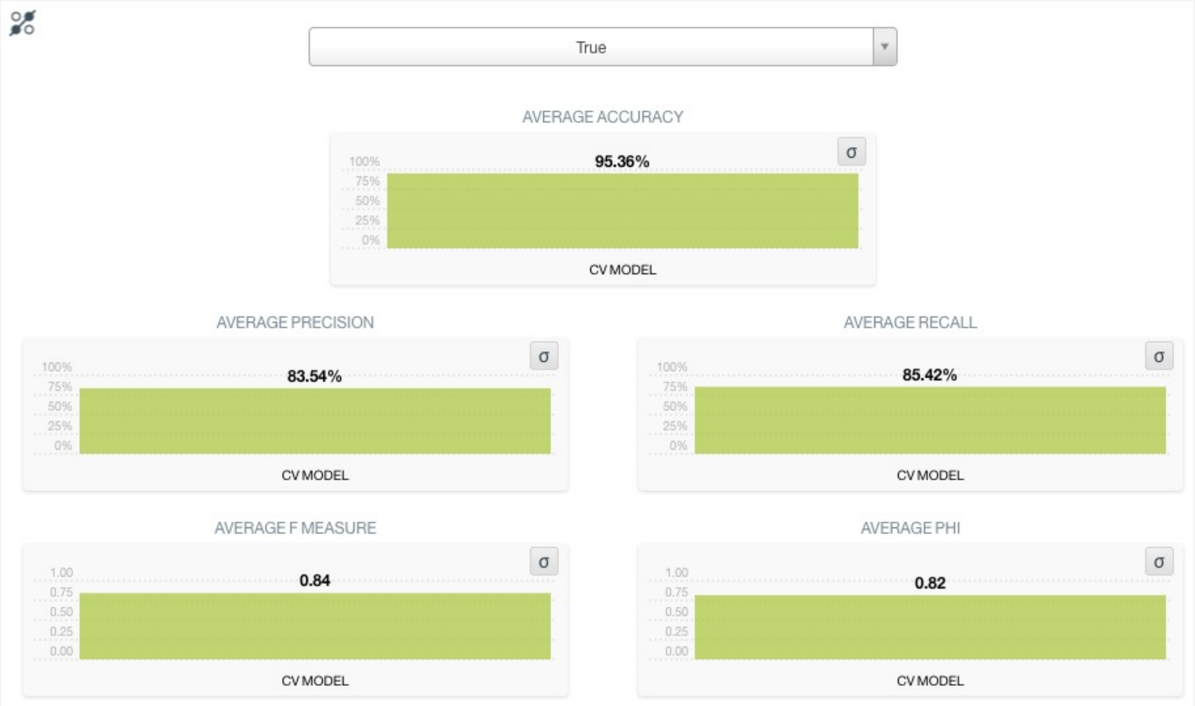
5-fold cross validation configuration



Individual evaluations

- **5-fold evaluation Dataset-Assignment1**
  512-node, pruned, deterministic order, 4 datasets, o…
- **4-fold evaluation Dataset-Assignment1**
  512-node, pruned, deterministic order, 4 datasets, o…
- **3-fold evaluation Dataset-Assignment1**
  512-node, pruned, deterministic order, 4 datasets, o…
- **2-fold evaluation Dataset-Assignment1**
  512-node, pruned, deterministic order, 4 datasets, o…
- **1-fold evaluation Dataset-Assignment1**
  512-node, pruned, deterministic order, 4 datasets, o…

5-fold cross validation average results



esade

Appendix: Model finetuning

13 Fine-tune your models and try to improve their performance.

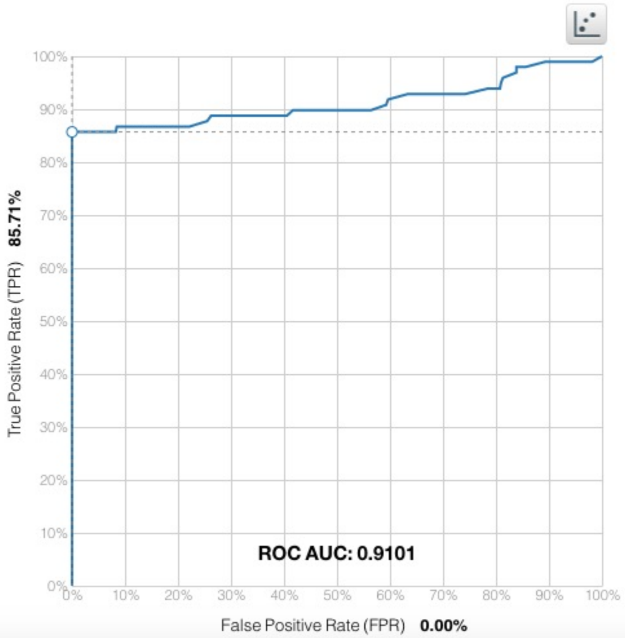
! *Technical details on Advanced Hyperparameters on answer slide*

Confusion Matrix and Performance Metric of 9th model (Trial 9) - Decision Tree

Trial 9

ACTUAL VS. PREDICTED	True	False	ACTUAL	RECALL
True	84	14	98	85.71%
False	0	545	545	100.00%
PREDICTED	84	559	643	92.86% AVG. RECALL
PRECISION	100.00%	97.50%	98.75% AVG. PRECISION	97.82% ACCURACY

	F	Phi
True	0.92	0.91
False	0.99	0.91
	0.96 AVG. F	0.91 AVG. Phi



97.8% Accuracy	0.9231 F-measure	
100.0% Precision	85.7% Recall	0.9142 Phi coefficient
0.0% FPR	13.1% % positive instances	656.1% Lift
85.7% K-S statistic	0.4177 Kendall's Tau	0.5106 Spearman's Rho

Appendix: Evaluating model against test dataset

14 Are your models overfitted? How did you check this? Why is an overfitted model useless?

Confusion Matrix of model trail 9 against test set

Positive class: True

ACTUAL VS. PREDICTED	False	True	ACTUAL	RECALL	F	Phi
False	545	0	545	100.00%	0.99	0.91
True	14	84	98	85.71%	0.92	0.91
PREDICTED	559	84	643	92.86% AVG. RECALL	0.96 AVG. F	0.91 AVG. Phi
PRECISION	97.50%	100.00%	98.75% AVG. PRECISION	97.82% ACCURACY		

97.8% Accuracy	0.9231 F-measure

100.0% Precision	85.7% Recall	0.9142 Phi coefficient

Confusion Matrix of model trail 9 against training set

Positive class: True

ACTUAL VS. PREDICTED	False	True	ACTUAL	RECALL	F	Phi
False	2,195	6	2,201	99.73%	0.99	0.90
True	54	314	368	85.33%	0.91	0.90
PREDICTED	2,249	320	2,569	92.53% AVG. RECALL	0.95 AVG. F	0.90 AVG. Phi
PRECISION	97.60%	98.13%	97.86% AVG. PRECISION	97.66% ACCURACY		

97.7% Accuracy	0.9128 F-measure

98.1% Precision	85.3% Recall	0.9023 Phi coefficient

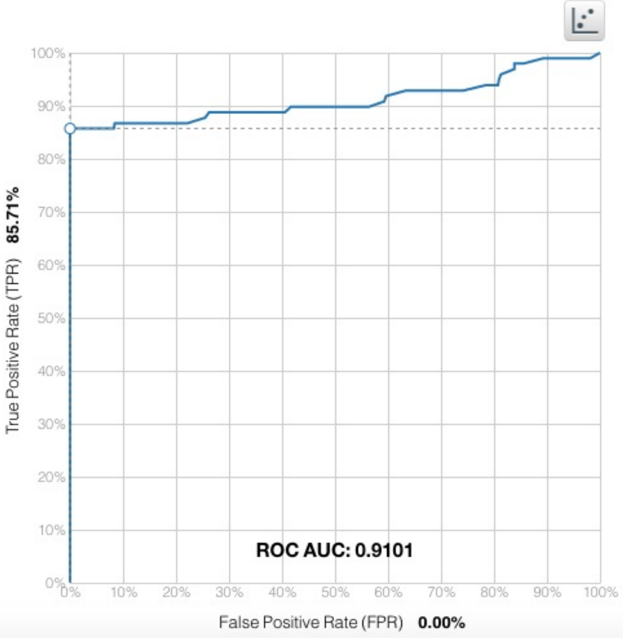
Appendix: Confusion Matrix Interpretation of Best Model

16 Interpret the Confusion Matrix of one of the models. What represents each metric (Accuracy, Recall...) and output (TP, TN...) in technical and business terms?

Confusion Matrix and Performance Metric of 9th model (Trial 9) - Decision Tree

Trial 9

ACTUAL VS. PREDICTED	True	False	ACTUAL	RECALL	F	Phi
True	84	14	98	85.71%	0.92	0.91
False	0	545	545	100.00%	0.99	0.91
PREDICTED	84	559	643	92.86% AVG. RECALL	0.96 AVG. F	0.91 AVG. Phi
PRECISION	100.00%	97.50%	98.75% AVG. PRECISION	97.82% ACCURACY		



97.8% Accuracy	0.9231 F-measure	
100.0% Precision	85.7% Recall	0.9142 Phi coefficient
0.0% FPR	13.1% % positive instances	656.1% Lift
85.7% K-S statistic	0.4177 Kendall's Tau	0.5106 Spearman's Rho

Best model

esade

Thank you

Matteo Giardini

esade