Management Summary

BoBo, established in 2014, is a French B2B food catering company that offers meal services for companies and company events.

Situation: In April 2020, Kunstvoll, a competitor from Germany wants to start offering a similar subscription service in France, targeting small to mid-size companies, with prices on average 15% lower than BoBo.

Project Goal: Estimate the churn risk for various churn cutoff points (e.g., 5%, 10%, 15%). Provide BoBo a list of customers with a high churn risk.

1. Analysis

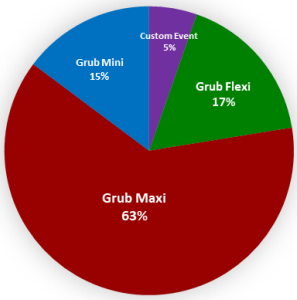
We performed a deep dive on the data provided and on our results to understand BoBo's current and future situation. After creating our Basetable and running our Predictive Model, we analyzed the variables with the highest score according to the Random Forest Feature Importance (see Appendix 4).

We analyzed Normal and Average Deliveries, and concluded that customers who have less than 10 deliveries have a high churning incidence, of approximately 30%. In addition, we analyzed the incidence churn by subscriptions' price. (see Appendix 1)

2. Results – Active Customers

Our results identified 860 active customers (customers who renewed their subscriptions before February 2019). These customers had an average subscription price of €3,898 and, on average, six total subscriptions.

The bulk of these customers are located in two regions, 57% are located in region five and 38% in region one. Graph 1 shows the product distribution of the 860 active customers in their last subscription.



Graph 1. Distribution of Active Customers' Last Subscription

3. Results – Churners

Probability	Nbr. Clients
≥ 95%	1
≥ 90%	1
≥ 85%	1
≥ 80%	3
≥ 75%	101
≥ 70%	140
≥ 65%	191
≥ 60%	223
≥ 55%	247

Table 1. Churners at various cutoff points

We used a Random Forest Classifier to predict the churn risk of active customers. Table 1 shows the cumulative number of customers our model predicts will churn, at various cutoff points. We have highlighted the most significant cutoff points.

The results obtained predict 247 customers will churn; this represents 29% active customers, and possible loss of €4,688 per customer, on average.

We have compiled a list of 100 customers with a 75% or higher churn risk. This list, with the CustomerID, Region, Average Total Sales, Total Subscriptions, Total Renewals, and Last Subscription Name can be found on the MS Excel file named *Top_100_Churners*. A shorter version, containing only CustomerID, is shown in Appendix 2 – Top 100 Churners.

Technical Summary

BoBo – Food Catering Company – provided five databases: 1. Customers: List of every customer the company has had, including current ones. 2. Subscriptions: List of the subscriptions that customers have acquired and their conditions, 3. Complaints: Complaints made by customers since February 2011, 4. Delivery: Information about deliveries made to customers, and 5. Formula: It gives us information about how offers where made to customers, through email or regular intake. Finally, Spark – Databricks was utilized for the development of this project.

1. Data Summary and Processing

Data Summary

- **Variables:**

The target variable in this project is the churn response, which refers to a customer leaving the company or not. In the case of BoBo Company, churners refer to the customers who will not renew their subscriptions. The response was marked with 0 and 1 (Binary Variable). If the customer does not renew their subscription (churner) the response variable is equal to 1, otherwise it is equal to 0.

The Predictor Variables are shown in Appendix 3; however, for building the different models we performed Feature Selection for decreasing the number of Predictor Variables.

- **Timeline:**

The information provided by BoBo has the following Time Range: from February 2011 for Complaints, January 2014 for Deliveries and Subscriptions, until February 2019 for the three databases. We used information from January 2014 to January 2018 to create our Basetable and training the Predictive Model.

For our Training Predictive Model, the timeline for the "Past Snapshot" of the Independent Variables (customer information) is from January 2014 to January 2018. The Dependent Variable was evaluated on the following basis: if the customer had renewed their subscription before January 2018 the target variable was labeled "0" (Not Churner), if the customer had not renewed their subscription the target variable was labeled "1" (Churner).

Our Final Predictive Model was applied to the time range February 2019 – *Present.* It was applied to the current customers to determine the probability of Churn – the probability that they will not renew their current subscription.

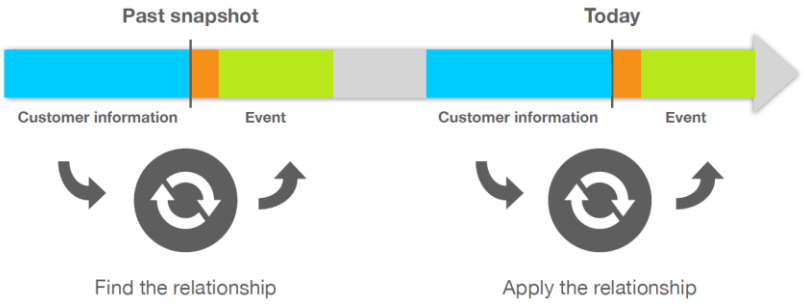


Figure 1. Timeline

- Data Visualization:

The following figure (Figure 2) shows how our Target Variable is distributed in the Basetable. There is a churning incidence of 25.4% (342 targets).

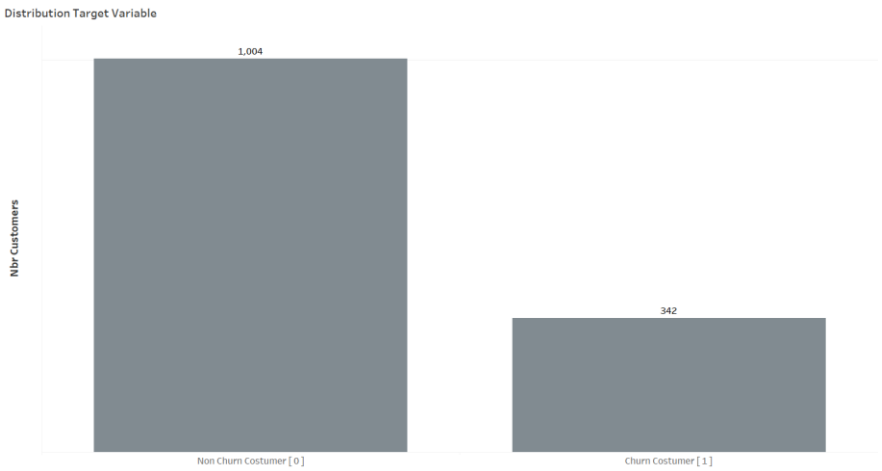


Figure 2. Distribution target variable in the base table; 1: Churn Customer, 0: Non Churn Customers

Preprocessing Data

For the tables Complaints, Subscriptions, and Delivery, we made aggregated calculations grouped by CustomerID. For the Delivery table, we matched the CustomerID using the SubscriptionID in the Subscriptions table – which contains both, the CustomerID and the SubscriptionID.

We then merged the three tables by CustomerID. Additionally, we merged the Formula table by FormulaID with the other three tables (Complaints, Subscriptions, and Delivery). Finally, we merged all of them to the Customers table to have our final Basetable.

Regarding missing data, we dropped subscriptions that contained missing values from columns GrossFormulaPrice to TotalCredit. Additionally, we verified that these subscriptions did not have any deliveries in the Delivery table.

We changed the Product Name in order to delete the period it contained from original information (e.g. “Grub Flexi (excl. staff)”).

The final base table contains information about 1.347 customers. It has 58 columns (see appendix 1 – Variable Description) and each row belongs to a different customer (unique CustomerID).

Feature Engineering

We created 43 variables (see appendix 3 – Variable Description) taking into account the different information provided in the tables. Other variables were taken from the last subscription of each customer (subscription before January 2018).

Variable Selection

We used Chi Square Selector method for reducing the number of variables and considering only the ones that are not dependent on the variable response. We selected the best 15 variables and they were included in each model that we trained.

2. Methodology

Methods

BoBo's situation involves classifying their customers as Churners (label=1) or Non-Churners (label=0). Thus, we have to use Classifier Algorithms for training the Predictive Model. We chose to use the following algorithms in the phase model: Random Forest Classifier, and Logistic Regression.

- **Random Forest Classifier:** We trained the model tuning the parameter numTrees with the values [150, 300, 500].
- **Logistic Regression:** We trained the model tuning the parameter regParam with the values [0.1, 0.01].

Experimental Setup

The Basetable was randomly divided into train and test set, and stratified by the label to have the same proportion of the Target Variable in both sets. The proportion used was 0.8 (train) and 0.2 (test). This method works well because all observations are independent; the information of one customer does not affect the information of another customer. The train set has 1,075 observations, and the test set has 271 observations.

In all cases, for training the model, we performed cross-validation with two folds. The metric used in cross-validation for choosing the best model was AUC.

Additionally, we standardized the numerical variables (mean=0, standard deviation=1). We also created dummy variables for the variables ProductName_last, FormulaType_last, PaymentStatus_last and Region.

3. Results

Model Results and Interpretation

After training the models, we tested them on the Test Set to evaluate their performance in unseen data. The metrics used to compare the models are ROC AUC and PR AUC. Higher values on these metrics signal better performance. The results are showed in Table 2.

Model	ROC AUC	PR AUC
Random Forest Classifier	0.8325	0.7824
Logistic Regression	0.8418	0.7696

Table 2. Benchmark performance Random Forest Classifier and Logistic Regression.

After the benchmark of the models, we selected Random Forest Classifier because of its better performance in PR AUC: 0.7824. The ROC AUC for both models is very similar, thus we focused on PR AUC and interpretability when selecting our model. Additionally, we performed a Feature Importance (see Appendix 4) in the Random Forest Predictive Model.

Prediction current customers

Using the selected model, we apply it to BoBo's current customers for predicting what customers have the highest probability to churn. Current customers are selected based on their last renewed subscription (renewed subscription before February 2019).

Appendix 1 – Normal Deliveries, Total avg. Deliveries and Total Price Subscription for BoBo's customers

Appendix 1. Distribution of BoBo's Customers Deliveries and Churning Incidence – Approximately 30% Churning Incidence

In the previous graph, we can observe that clients who pay less for their subscription are the ones who have greater incidence rate, approx. 0.5. It can be suggested that this rate could be greater in this group of clients if they have the opportunity to find a lower price.

Appendix 2 – Top 100 Churners

Customer ID									
1173686	109862	793482	784650	97924	217062	69926	238227	775416	765834
1119917	72290	198700	792661	126371	460459	745470	501664	38731	672514
1012153	135532	658311	405673	88880	190317	671074	119412	80863	851071
296777	92478	129707	494173	83829	274943	493334	123310	1104132	688993
591609	119308	69031	132752	126990	744773	78403	62360	105447	664959
294578	658395	487392	42801	108424	215157	103873	95777	124783	667763
168511	71104	100633	127178	101454	148606	123674	122454	71100	653360
113134	85354	752673	689119	282163	59580	121120	43291	768448	682856
131727	90987	665396	80340	256376	216874	65777	112472	101481	761505
774120	114210	685953	85713	797994	139859	736906	187487	141979	959735

Appendix 2. List of 100 CustomerID with a 75% or higher churn risk.

Appendix 3 – Variable Description Base Table

Variable Name	Description
CustomerID	Unique numeric identifier per Customer
Region	Anonymized numeric identifier of the customer's Region
StreetID	Anonymized numeric identifier of the customer's Street
total_subscriptions	Number of total subscriptions per Customer
date_last_Endsubs	Customer's last subscription end date
avg_duration_sub	Average subscription duration in months
total_renewals	Sum of subscription renewal per customer
total_credits	Total credits per customer
avg_mealsREG	Average number of meals to be provided on the regular basis
avg_mealsEXCEP	Average number of meals to be provided on exceptional basis
avg_form_price	Average price of customer's subscriptions
avg_price_meal	Average price per meal
avg_product_dcto	Average discount on product
avg_formula_dcto	Average discount on formula
avg_total_price	Average price paid by customer
RenewalDate_last	Customer's last subscription renewal date
StartDate_last	Start date of last subscription
ProductName_last	Customer's last subscription product name
FormulaType_last	Customer's last subscription formula type (CAM = intake through a direct mail campaign; REG = regular intake)
duration_last	Customer's last formula duration in months
PaymentStatus_last	Customer's last subscription payment status
max_SartDate	Customer's last subscription start date
nbrProd_custom_events	Number of times a customer has had a Custom Event product
nbrProd_custom_events	Number of times a customer has had a Flexi product
nbrProd_custom_events	Number of times a customer has had a Maxi product
nbrProd_custom_events	Number of times a customer has had a Mini product
recency_endSubs	Number days since last subscription
missing_delivery_class	Number of times a customer's delivery was missing
nbr_abnormal_deliveries	Number of times a customer had an abnormal delivery
nbr_normal_deliveries	Number of times a customer had a normal delivery
total_deliveries	Total deliveries per customer
last_delivery	Last delivery date
avg_days_inter_del	Average number of days between deliveries per customer
avg_nbr_deliveries_subs	Average number of deliveries per customer
nbrComp_billing_incorrect	Number of times a customer complained about incorrect billing
nbrComp_billing_incorrect	Number of times a customer complained about rude employees
nbrComp_billing_incorrect	Number of times a customer complained about food quality
nbrComp_billing_incorrect	Number of times a customer complained about food quantity
nbrComp_billing_incorrect	Number of times a customer complained about food being cold
nbrComp_billing_incorrect	Number of times a customer complained about delivery lateness
nbrComp_billing_incorrect	Number of times a customer complained about incorrect order
nbrComp_billing_incorrect	Number of times a customer complained about a miscellaneous complaint
nbrComp_billing_incorrect	Number of times a customer complained about food hygiene

total_complaints	Total number of Complaints per customer
last_complaint	Date of last complaint
nbrSln_na_solution	Number of times a NA solution was offered to a customer
nbrSln_na_solution	Number of times a Price Discount solution was provided to a customer
nbrSln_na_solution	Number of times an Additional Meal solution was provided to a customer
nbrSln_na_solution	Number of times a No Compensation solution was provided to a customer
nbrSln_na_solution	Number of times a Miscellaneous solution was provided to a customer
na_feedback	Number of times a customer provided NA as feedback for the solution offered
no_ans_feedback	Number of times a customer provided No Answer as feedback for the solution offered
no_satisfied	Number of times a customer was Not Satisfied with the solution provided
other_feedback	Number of times a customer provided Other as feedback for the solution offered
satisfied	Number of times a customer was Satisfied with the solution provided
rate_sln	Percentage a customer was offered a Solution
rate_no_satf	Percentage a customer was Not Satisfied with the Solution
label	Churn Label, 1:Churn; 0:No Churn

Appendix 4 – Random Forest Feature Importance

	idx	name	score
2	2	total_renewals	0.260747
13	13	nbr_normal_deliveries	0.203076
14	14	total_deliveries	0.117855
0	0	total_subscriptions	0.085963
11	11	recency_endSubs	0.068472
8	8	avg_total_price	0.067743
5	5	avg_form_price	0.042415
6	6	avg_price_meal	0.035622
7	7	avg_formula_dcto	0.033839
4	4	avg_mealsEXCEP	0.022581
3	3	avg_mealsREG	0.018902
1	1	avg_duration_sub	0.014141
12	12	nbr_abnormal_deliveries	0.010912
9	9	duration_last	0.010773
10	10	nbrProd_maxi	0.006959

Appendix 4. Random Forest Feature Importance – Top 15 most important variables for Final Predictive Model