Customer Behavior Prediction

Giovanni Fossati

GEICO – 2015.09.06

Summary

GEICO policy holders have access to different ways to make payments, comprising several self-service options (e.g. webpage, mobile Apps, automated phone system).

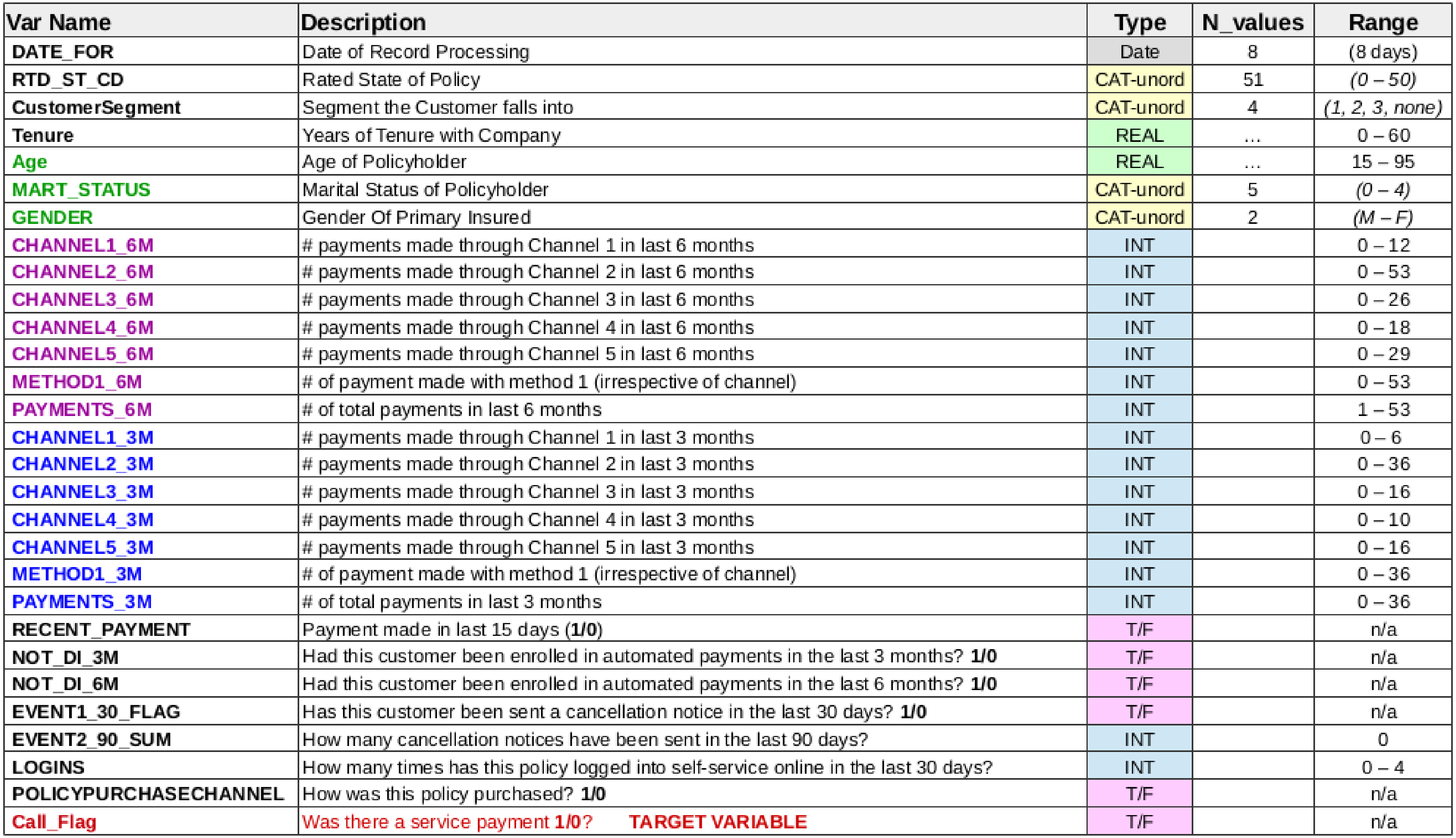
A small fraction of the customers makes payments over the phone with the assistance of a service counselor, a channel that carries higher cost for the Company and for the customer.

- The **goal** of the analysis presented here is to identify customers who are likely to make a *service payment call* and to send them targeted emails close to the time when their payments are due to encourage them to pay via one of the self-service channels.
- The training **dataset** includes information about "demographics" and "payment behavior" for about 130,000 customers, of whom around 3.7% made a service payment call.
- Small fractions like this present a challenge for predictive modeling, even when the size of the data set is large. A "dummy" model predicting 'FALSE' (no call) for all customers would be 96.3% accurate, setting the bar very high, while being completely useless.
- So, beside careful with the selection of a method, one has to be thoughtful about the ***choice of the measure of quality of the model performance***.
- We modeled a few different data set, comprising different mix of original and new variables, with two different methods: **Logistic Regression** and **Multivariate Adaptive Regression Splines** (MARS). The analysis and modeling was performed with **R**.
- After evaluating each method on different sets of variables, on the basis of Sensitivity and Specificity, we deemed the MARS results to provide better predictive power. The inclusion of a few new variables summarizing (grouping) some of the numeric variables seems to add discriminatory ability to the model.

The Data

The data set comprises 28 variables plus the outcome (`Call_Flag`), for 130086 policyholders.

- 4 are unordered categorical variables.
- Of the numeric variables, 2 are real, 14 are integers.
- 6 are binary (including the outcome).



Var Name	Description	Type	N_values	Range
DATE_FOR	Date of Record Processing	Date	8	(8 days)
RTD_ST_CD	Rated State of Policy	CAT-unord	51	*(0 – 50)*
CustomerSegment	Segment the Customer falls into	CAT-unord	4	*(1, 2, 3, none)*
Tenure	Years of Tenure with Company	REAL	…	0 – 60
Age	Age of Policyholder	REAL	…	15 – 95
MART_STATUS	Marital Status of Policyholder	CAT-unord	5	*(0 – 4)*
GENDER	Gender Of Primary Insured	CAT-unord	2	*(M – F)*
CHANNEL1_6M	# payments made through Channel 1 in last 6 months	INT		0 – 12
CHANNEL2_6M	# payments made through Channel 2 in last 6 months	INT		0 – 53
CHANNEL3_6M	# payments made through Channel 3 in last 6 months	INT		0 – 26
CHANNEL4_6M	# payments made through Channel 4 in last 6 months	INT		0 – 18
CHANNEL5_6M	# payments made through Channel 5 in last 6 months	INT		0 – 29
METHOD1_6M	# of payment made with method 1 (irrespective of channel)	INT		0 – 53
PAYMENTS_6M	# of total payments in last 6 months	INT		1 – 53
CHANNEL1_3M	# payments made through Channel 1 in last 3 months	INT		0 – 6
CHANNEL2_3M	# payments made through Channel 2 in last 3 months	INT		0 – 36
CHANNEL3_3M	# payments made through Channel 3 in last 3 months	INT		0 – 16
CHANNEL4_3M	# payments made through Channel 4 in last 3 months	INT		0 – 10
CHANNEL5_3M	# payments made through Channel 5 in last 3 months	INT		0 – 16
METHOD1_3M	# of payment made with method 1 (irrespective of channel)	INT		0 – 36
PAYMENTS_3M	# of total payments in last 3 months	INT		0 – 36
RECENT_PAYMENT	Payment made in last 15 days **(1/0)**	T/F		n/a
NOT_DI_3M	Had this customer been enrolled in automated payments in the last 3 months? **1/0**	T/F		n/a
NOT_DI_6M	Had this customer been enrolled in automated payments in the last 6 months? **1/0**	T/F		n/a
EVENT1_30_FLAG	Has this customer been sent a cancellation notice in the last 30 days? **1/0**	T/F		n/a
EVENT2_90_SUM	How many cancellation notices have been sent in the last 90 days?	INT		0
LOGINS	How many times has this policy logged into self-service online in the last 30 days?	INT		0 – 4
POLICYPURCHASECHANNEL	How was this policy purchased? **1/0**	T/F		n/a
Call_Flag	Was there a service payment **1/0**? **TARGET VARIABLE**	T/F		n/a

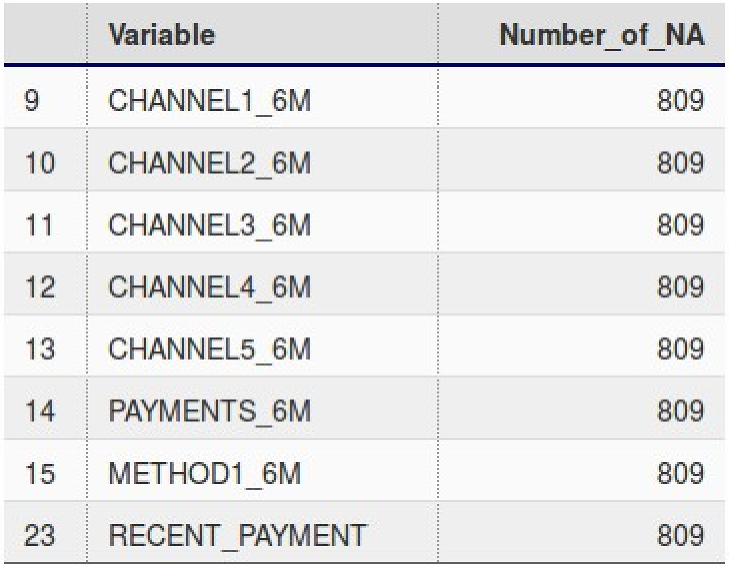
Exploratory Analysis – Missing Values

We reviewed a few important characteristics of the data, in particular focusing on

- Presence and distribution of **missing values** in the data set.
- The **distribution of the values** of the predictor.
- The **distribution of the outcome** as a function of the values of predictors.

Missing Values (NA)

- Eight variables have missing values, and they are all associated with payments in the last 6 months.
- In fact, all `NA` are the concentrated in the same 809 customers (0.6%), for whom there are not data for all their `*_6M` variables, as well as for `RECENT_PAYMENT`.
- They might be very new policyholders.
- A review of some of the characteristics of the customers with missing values does not seem to suggest that they are significantly different from the rest of the sample, for instance:
  - Their `Age` and `Tenure` are similar to those of the complete dataset, and do not support the “new customer” idea.
  - Their `GENDER` is a little skewed, but due to the small number this and other differences may simply be statistical fluctuations.
- Because of limited time at this stage we have not developed an imputation strategy.
  In the modeling we have either excluded them, or excluded the 8 variables with missing values.



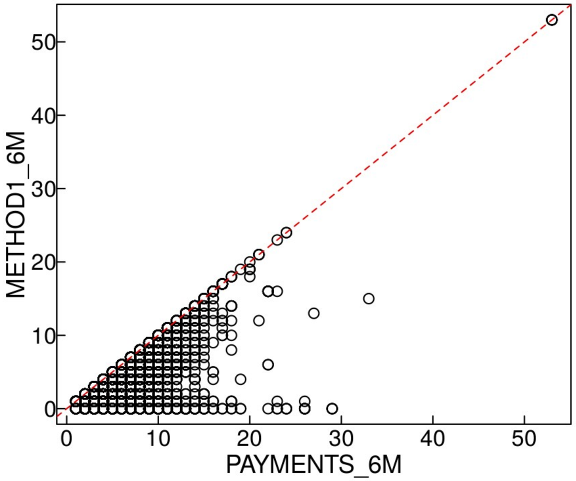
	Variable	Number_of_NA
9	CHANNEL1_6M	809
10	CHANNEL2_6M	809
11	CHANNEL3_6M	809
12	CHANNEL4_6M	809
13	CHANNEL5_6M	809
14	PAYMENTS_6M	809
15	METHOD1_6M	809
23	RECENT_PAYMENT	809

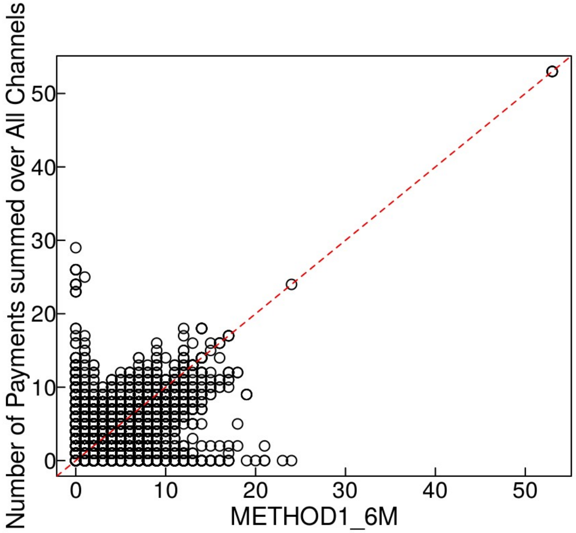
Exploratory Analysis – Imputation

- A simple strategy for imputation of missing values of variable (e.g. *CHANNEL1_6M*) is to *sample the distribution of the known values for customers with similar characteristics*.
  - A simple definition of similarity could be based on matching values of another relevant variable: for *CHANNEL1_6M* this latter could be *CHANNEL1_3M*.
  - We imputed missing values of **_6M* variables for a customer with **_3M = n* sampling the distribution of **_6M* for the subset of customers with **_3M = n*.
- The procedure goes as follows:
  - For each variable with missing values (e.g. *CHANNEL1_6M*) we prepared a matrix (contingency table) of the number of observations for each pair of values of *CHANNEL1_3M* and *CHANNEL1_6M*.
  - From this table we computed "probabilities" for the *CHANNEL1_6M* values at a fixed *CHANNEL1_3M* value by normalizing each row by the sum of the counts that row.
  - For each missing value of *CHANNEL1_6M*
    - we looked at the value of *CHANNEL1_3M* for that customer and
    - sampled a values of *CHANNEL1_6M* from the frequency distribution of *CHANNEL1_6M* values for customers with the same *CHANNEL1_3M.*
- For *RECENT_PAYMENT*, which is binary-valued, because for all customers with *RECENT_PAYMENT=NA* the PAYMENTS_3M=0, we can fill the missing values with 0s.

Exploratory Analysis - PAYMENTS

- One half of the predictors are related to mode of payment, and number of payments.
  - There are two groups of 7 variables each, summarizing number of payments over two different time frames: the last 6 and 3 months.
- It is not clear from the documentation what is the meaning of the different CHANNELS (5 of them), nor of METHOD. A count for payments made with METHOD1 (“irrespective of channel”) is given, and sometimes it is larger than the sum of payments over all channels.
  - Therefore CHANNELS and METHODS seem to refer to aspects of the paying process that are independent of each other and can be mixed in different combinations.
- The value of PAYMENTS variable is always larger than the sum of CHANNELS and of METHOD1, which is a good consistency check.
- There are customers with surprisingly large number of payments.
  - Without knowing more about payments policies and modes, it is difficult to assess on a meaningful basis whether or not large numbers are spurious, or genuine outliers.

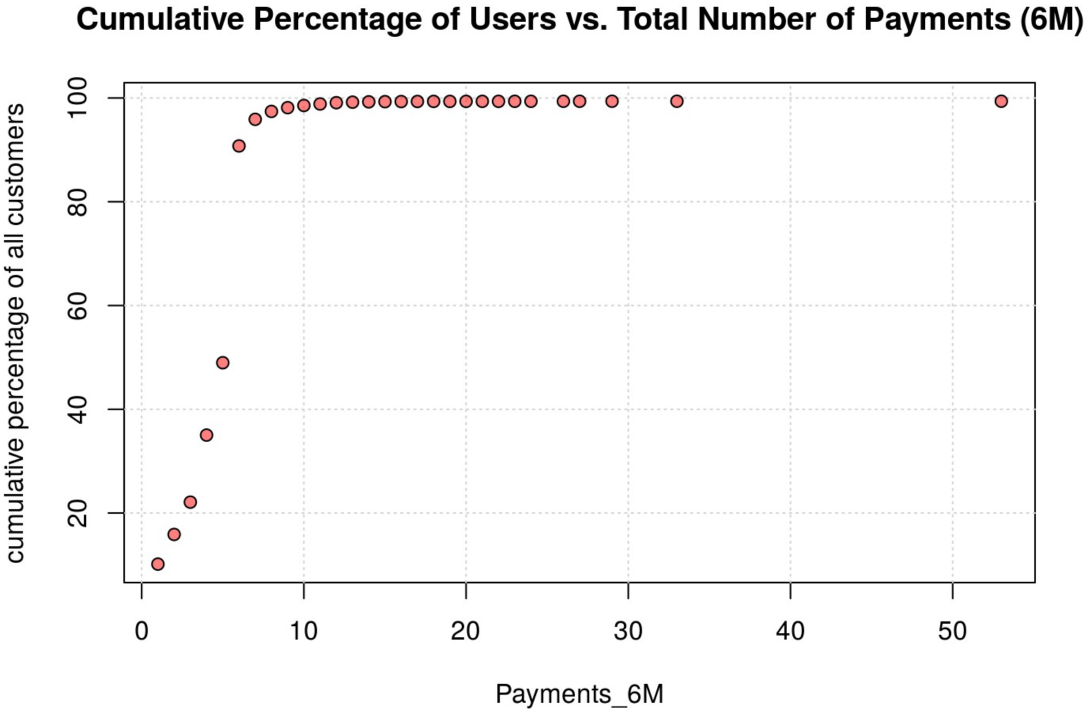




Exploratory Analysis – PAYMENTS cont.

- There are customers with surprisingly large number of payments.
  - Without knowing more about payments policies and modes, it is difficult to assess on a meaningful basis whether or not large numbers are spurious, or genuine outliers.
- The occurrence of extreme values is relatively rare.
  - Looking at PAYMENTS_6M, only 41 are >= 20 (0.027%), 690 >= 12 (0.5%).
  - Nevertheless, their pull on a model may be sufficient to impact its performance.

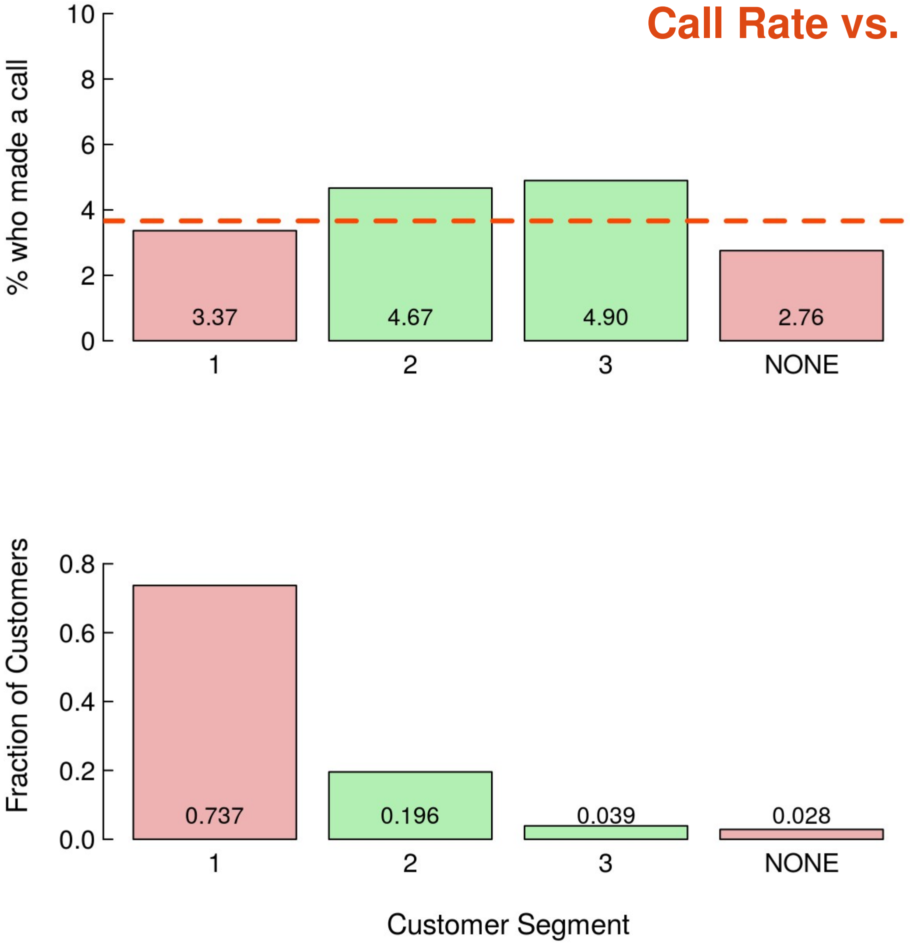
- At this stage we did not want to make a judgment call on the “nature” of the outliers and we tried a few approaches to deal with them: e.g.
  - transforming the variable to reduce their leverage (e.g. logarithm)
  - creating a new “tier” variable capturing the amount of customer activity in a more abstract form.
- The models we have tried, however do not seem to be very sensitive to the outlier, though this might be a sign of the inadequacy of these models for this task.

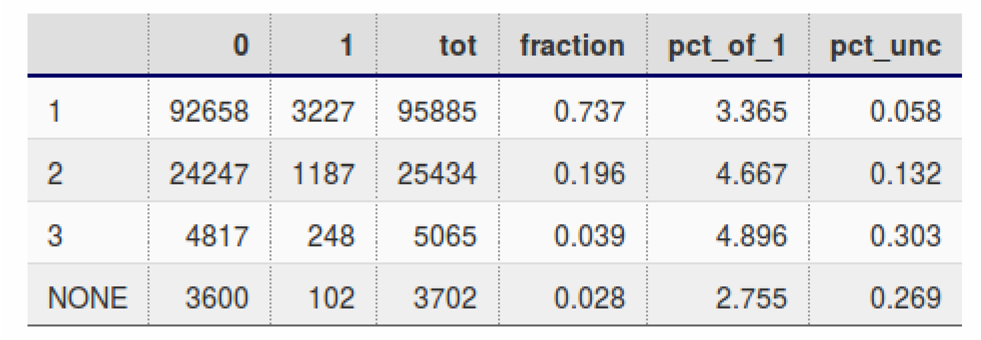


Exploratory Analysis – Outcome vs. predictors

- The average, baseline, “call rate” for the entire data set is around 3.66%.
- Computing the “call rate” by grouping the data according to the values of a predictor is a quick way to get some insight as to the importance of that predictor w.r.t. to its predictive power.
- We illustrate some examples in the following figures.
  - orange dashed line marks the base rate
  - bottom panels shows the fraction of the dataset in each group.)

Call Rate vs. Customer Segment



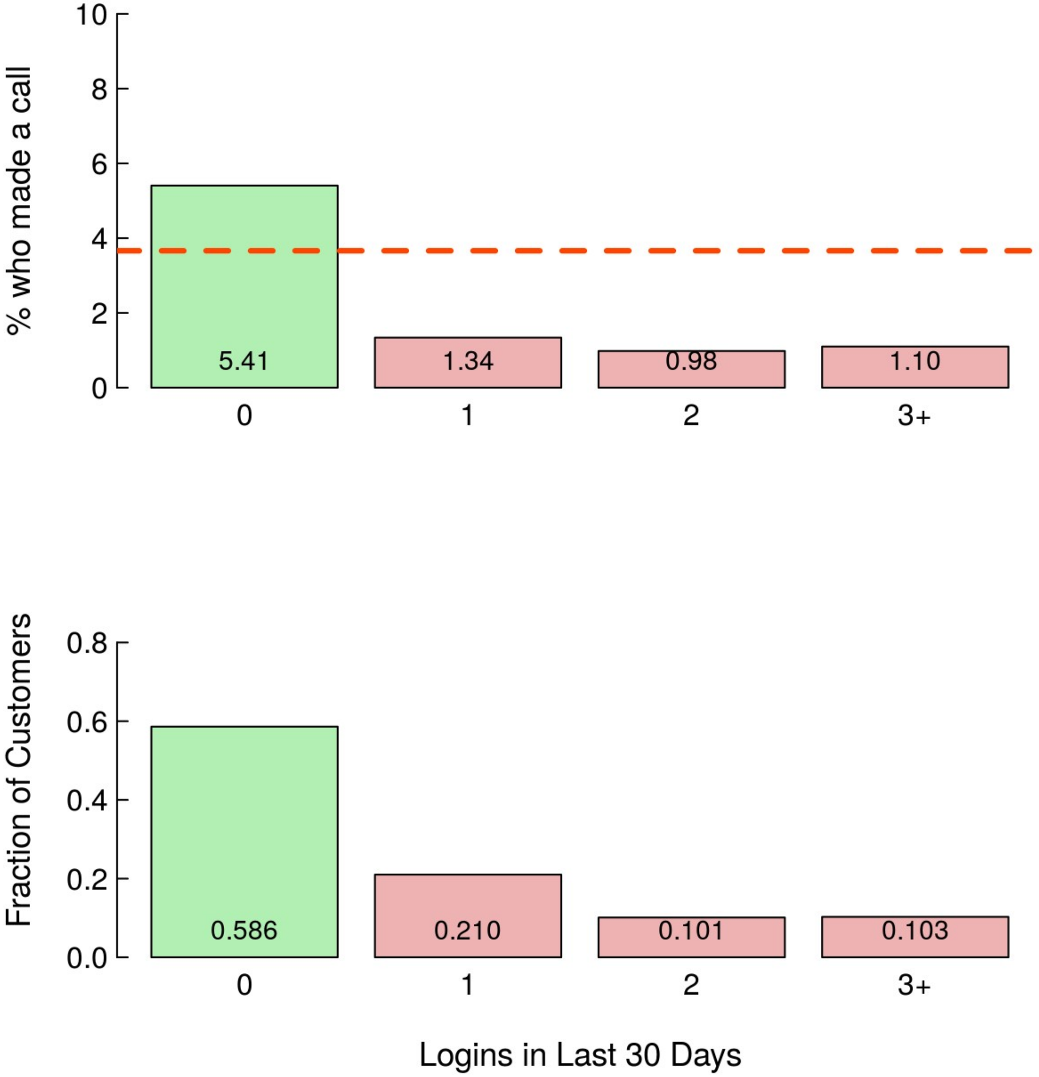


	0	1	tot	fraction	pct_of_1	pct_unc
1	92658	3227	95885	0.737	3.365	0.058
2	24247	1187	25434	0.196	4.667	0.132
3	4817	248	5065	0.039	4.896	0.303
NONE	3600	102	3702	0.028	2.755	0.269

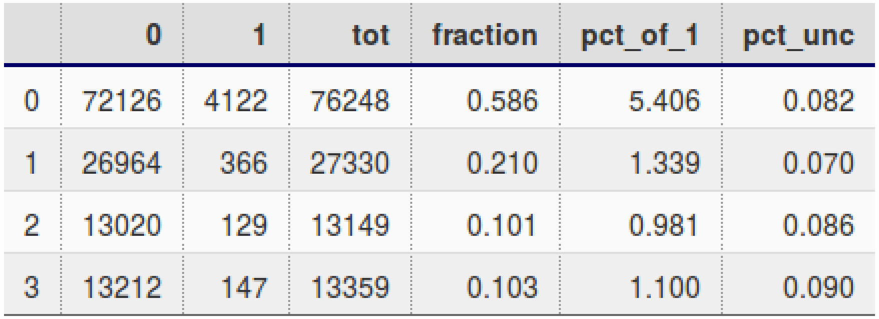
The variations in call rate between different customer segments are significant.

EDA – Outcome vs. LOGINS

Call Rate vs. Number of Logins in the last 30 days



Logins have been divided in four groups, with group 3+ including all customers with >= 3 logins.

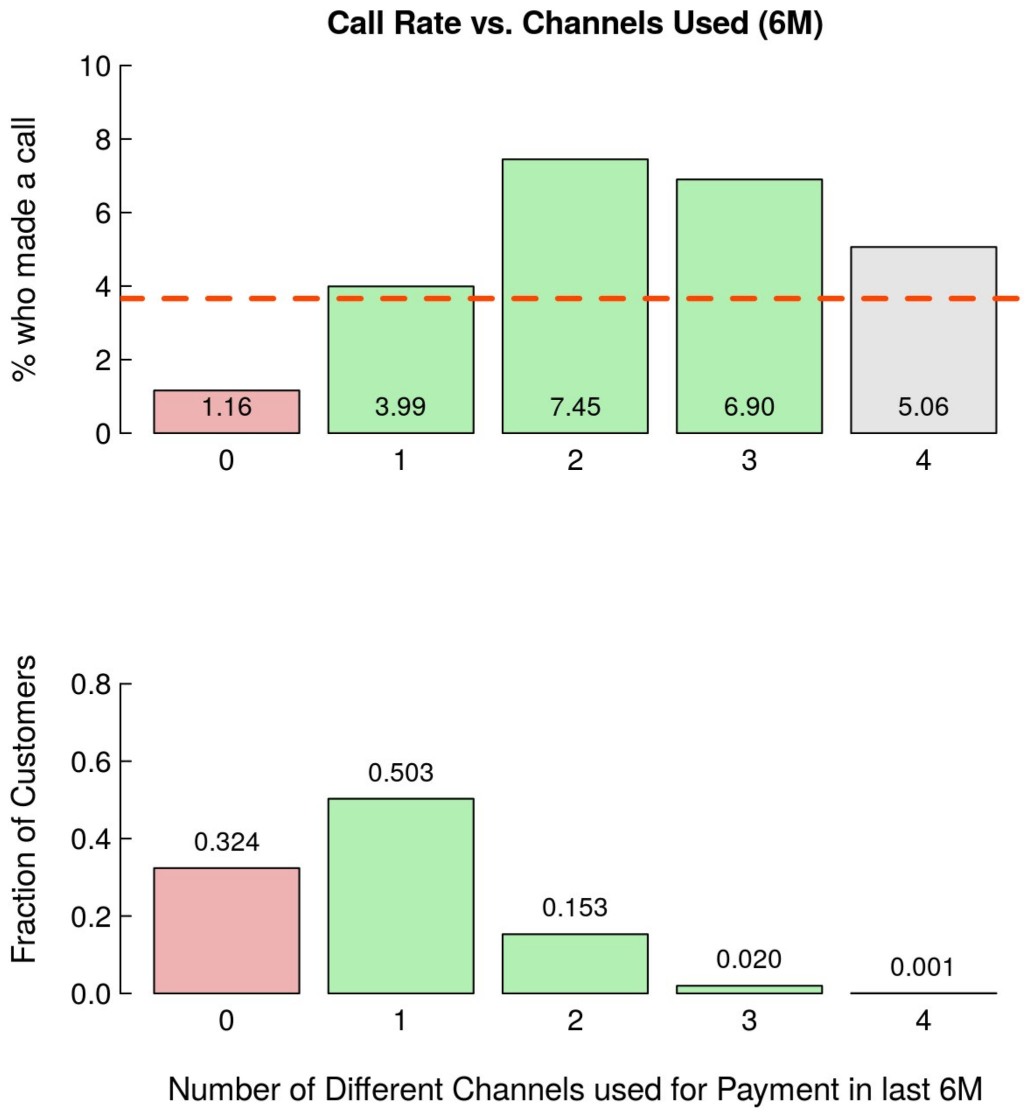


	0	1	tot	fraction	pct_of_1	pct_unc
0	72126	4122	76248	0.586	5.406	0.082
1	26964	366	27330	0.210	1.339	0.070
2	13020	129	13149	0.101	0.981	0.086
3	13212	147	13359	0.103	1.100	0.090

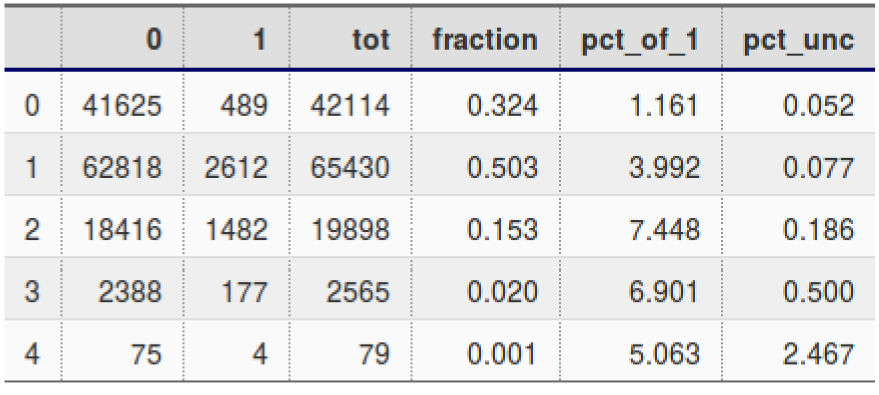
Perhaps not surprisingly customers who login regularly are significantly less likely to make a service payment call.

EDA – Outcome vs. *N of channels used*

Call Rate vs. Number of Different Channels Used



This variable counts the number of different channels used by a customer, irrespective of how many times each was used.



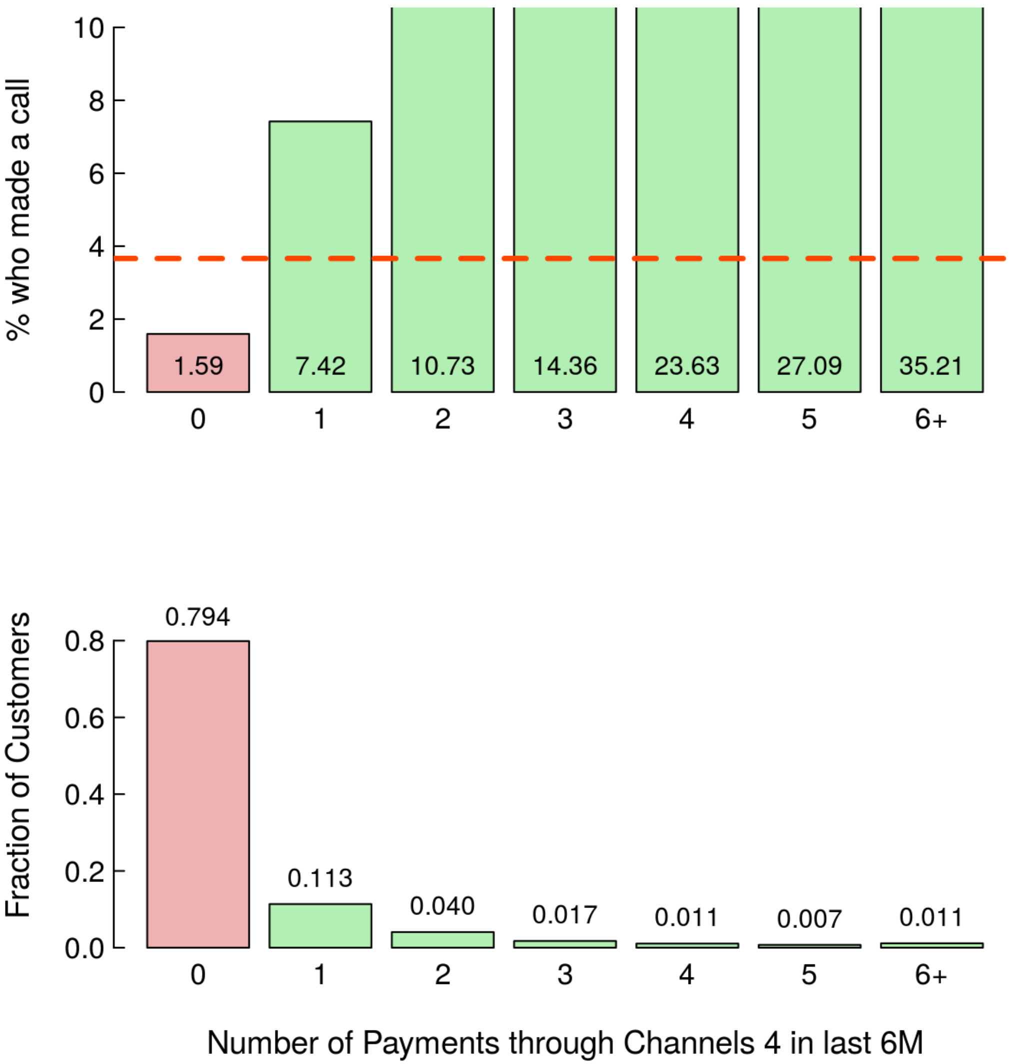
	0	1	tot	fraction	pct_of_1	pct_unc
0	41625	489	42114	0.324	1.161	0.052
1	62818	2612	65430	0.503	3.992	0.077
2	18416	1482	19898	0.153	7.448	0.186
3	2388	177	2565	0.020	6.901	0.500
4	75	4	79	0.001	5.063	2.467

Customers who use more channels are more likely to make a service call.

We do not have an intuitive interpretation of this effect, also because we do not know what the channels are.

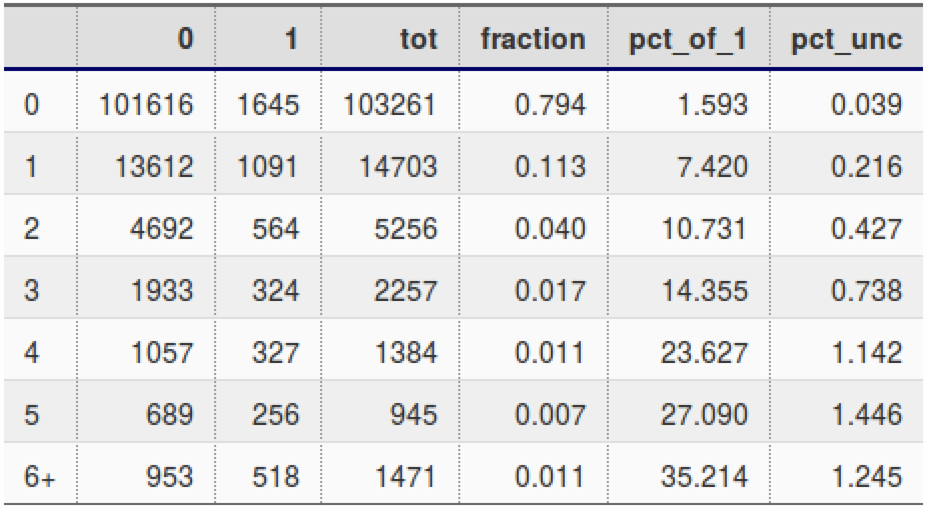
EDA – Outcome vs. *CHANNEL4_6M*

Call Rate vs. Number of Payments Made through Channel 4 in the last 6 Months



The `CHANNEL4_6M` variable emerged as one of the most important ones in all models we ran.

The effect on *Call Rate* is strikingly strong as illustrated by these plots and table.

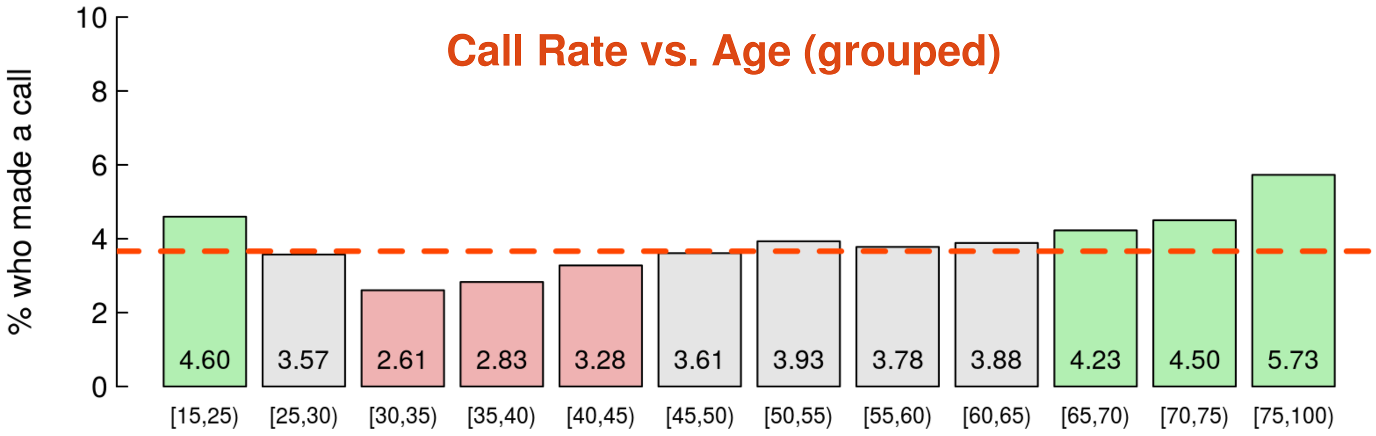


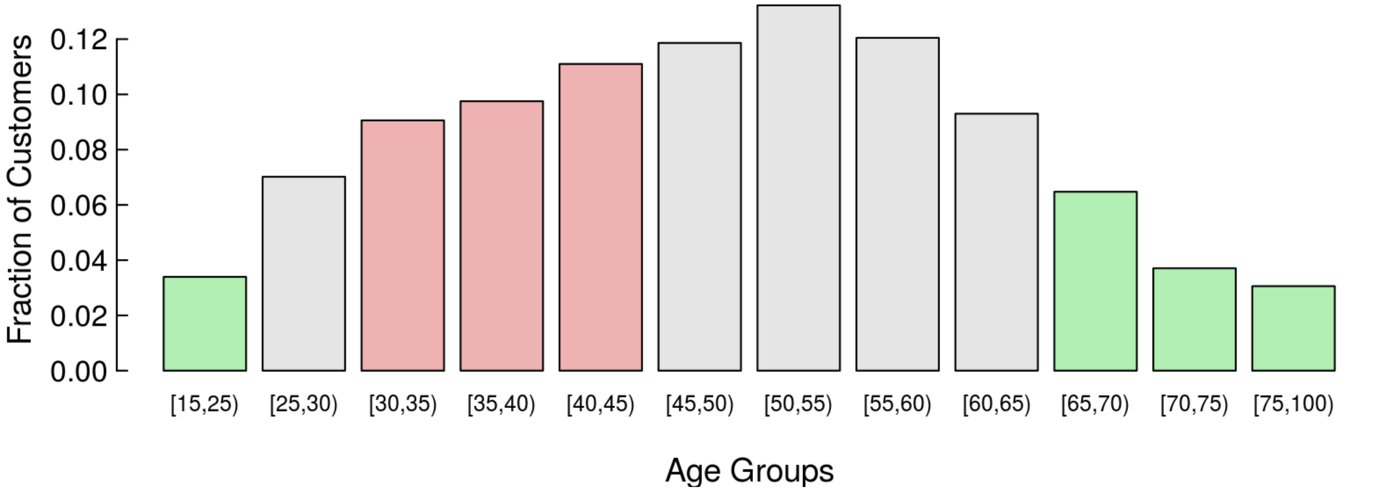
	0	1	tot	fraction	pct_of_1	pct_unc
0	101616	1645	103261	0.794	1.593	0.039
1	13612	1091	14703	0.113	7.420	0.216
2	4692	564	5256	0.040	10.731	0.427
3	1933	324	2257	0.017	14.355	0.738
4	1057	327	1384	0.011	23.627	1.142
5	689	256	945	0.007	27.090	1.446
6+	953	518	1471	0.011	35.214	1.245

Not knowing more about its meaning, namely what each CHANNEL means, we can not propose an interpretation for this effect.

One possible guess is that CHANNEL4 is itself a *phone-based payment channel* (e.g. automated phone system), hence showing the preference of a customer for phone vs. web-based interaction.

EDA – Outcome vs. `Age` (grouped)

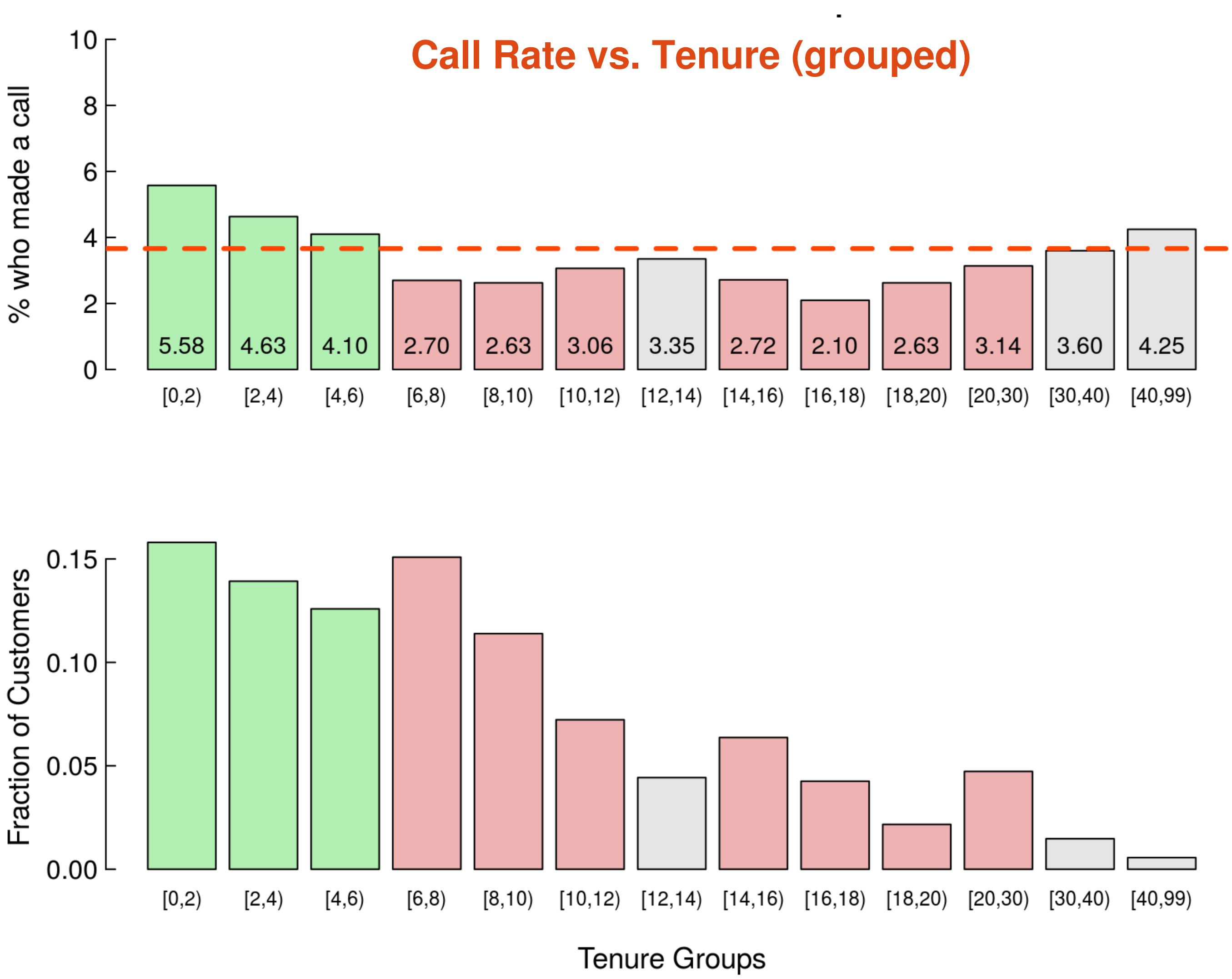




The “bathtub” shape of the variation of call rate with `Age` is significant.particular focusing on

- The higher rate at the low end can be interpreted as due to the *customer's inexperience.*
- The higher rate at the high end could be interpreted in terms of more *traditional* habits and inexperience with modern channels of communication of older customers.

EDA – Outcome vs. `Tenure` (grouped)



The “bathtub” shape of this plot (top panel) is closely related to the same causes of the shape of the relationship between call rate and age.

Modeling

We focused our effort on two fairly flexible methods.

- **Logistic Regression**, using R's *glm* function.
- **Multivariate Adaptive Regression Splines** (MARS), using R's *earth* package.
  - MARS offers an attractive combination of light computational “cost”, flexibility (thanks to its use of *hinge* functions), *built-in variable selection*, while conserving interpretatibility.
- For each method:
  - we trained models on datasets comprising different sets of variables,
  - split into training and testing subset for cross validation.

Evaluation Metrics

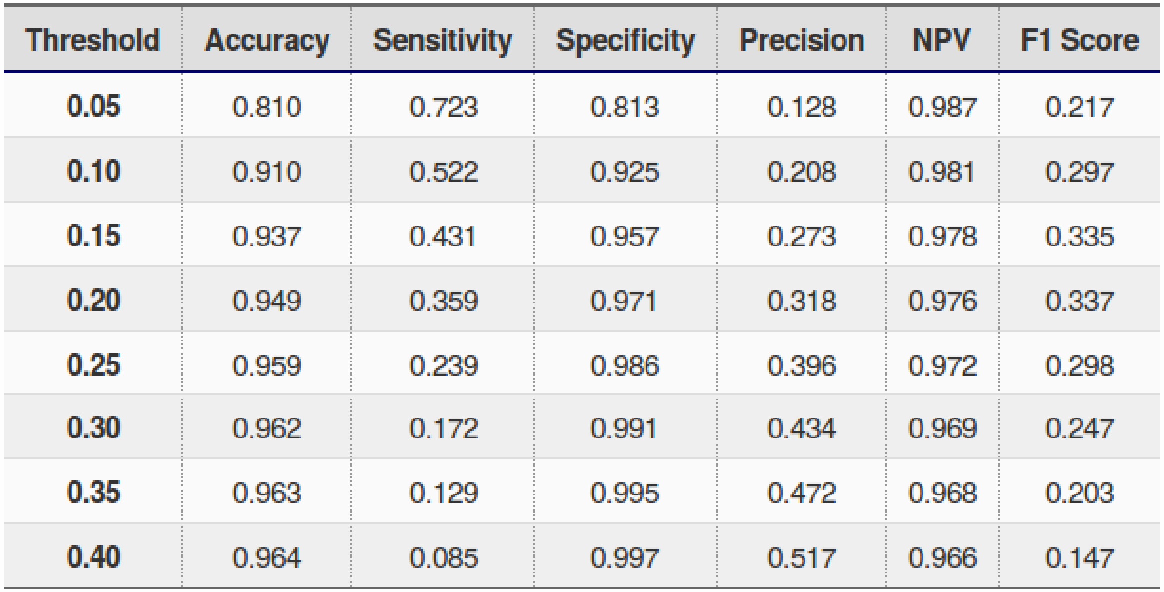
- The quality of each model (on each dataset) was assessed on the basis of metrics derived from *Confusion Matrices*, computed for different values of the threshold for assigning a class to the outcome variable.
  - For a quick and practical comparison of the results of the different models (method + variables set) we can rely on classic simple metrics such as *Accuracy*, *Sensitivity*, *Specificity* and *Area Under the Curve* (AUC).
  - However, as noted previously, the fact that the “desired” outcome is rare requires to take a more nuanced approach to evaluation, pondering the trade-off between *Accuracy*, *Sensitivity*, *Specificity*.

Model Evaluation – *Frequencies* and *Values*

- A more sophisticated approach would be to take into account the expected *value* (costs and benefits) of each combination of prediction and truth, combining the *Confusion Matrix* with a *Cost-Benefit Matrix*.
- The goal of the analysis is to identify users who may make a service call, which can be seen as a risk of incurring an avoidable cost (for GEICO), in order to avoid this by encouraging them to take advantage of the more cost-effective self-service channels.
  - One could think at the modeling task in the broader context of
    - the benefit of shifting a phone call to another (cheaper) channel and
    - the cost associated with the actions taken to encourage it (the emails).
  - By adding a probability of success of the effort then we would have the basic elements to combine the *Confusion Matrix* with a *Cost/Benefit Matrix* and compute an *expected value* for the model.
- Cast in this simplified context, the *expected value* is a function of two of the classic model performance measures: **sensitivity** and **specificity**.
  - *Increasing sensitivity* shifts users from False Negatives (customer is predicted to NOT call, but in reality she will) to True Positives: it increases the number of customers that we would target by email correctly, the customers from whom there could be a benefit.
  - Increasing sensitivity is accompanied by a *decrease in specificity,* i.e. more False Positives, customers who won't make a call but are predicted to do so. The increases of the number of “wasted” contacts has its own cost (no benefit).
- Depending on *Cost* (sending email), *Benefit* (savings of one spared phone call), *probability of “conversion”*, each model yields different expected values.

Model Evaluation – Adjusting *Threshold*

- **Sensitivity** and **specificity** are **changed by tuning the threshold** value at which we classify the outcome as YES/NO (1/0, T/F).
  - Changing threshold for a given Cost/Benefit Matrix yields different expected values.
- An example of the variation of key model performance metrics as a function of *classification threshold*, is shown in the following table (for case *MARS3*):

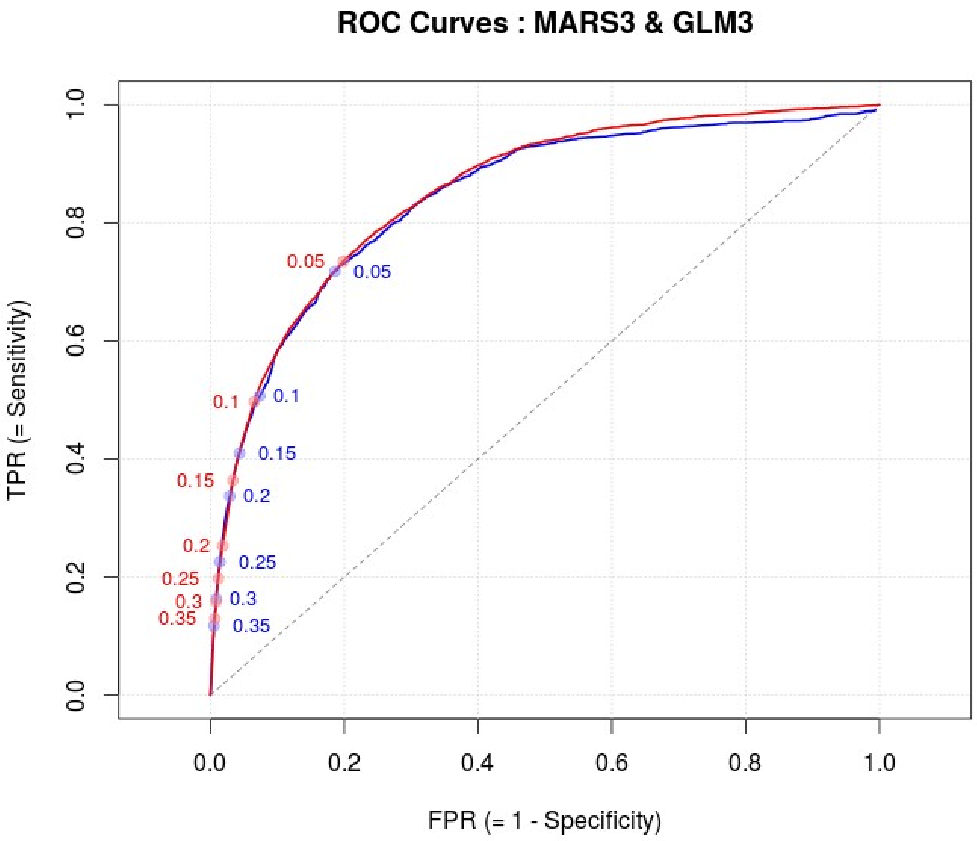


Threshold	Accuracy	Sensitivity	Specificity	Precision	NPV	F1 Score
0.05	0.810	0.723	0.813	0.128	0.987	0.217
0.10	0.910	0.522	0.925	0.208	0.981	0.297
0.15	0.937	0.431	0.957	0.273	0.978	0.335
0.20	0.949	0.359	0.971	0.318	0.976	0.337
0.25	0.959	0.239	0.986	0.396	0.972	0.298
0.30	0.962	0.172	0.991	0.434	0.969	0.247
0.35	0.963	0.129	0.995	0.472	0.968	0.203
0.40	0.964	0.085	0.997	0.517	0.966	0.147

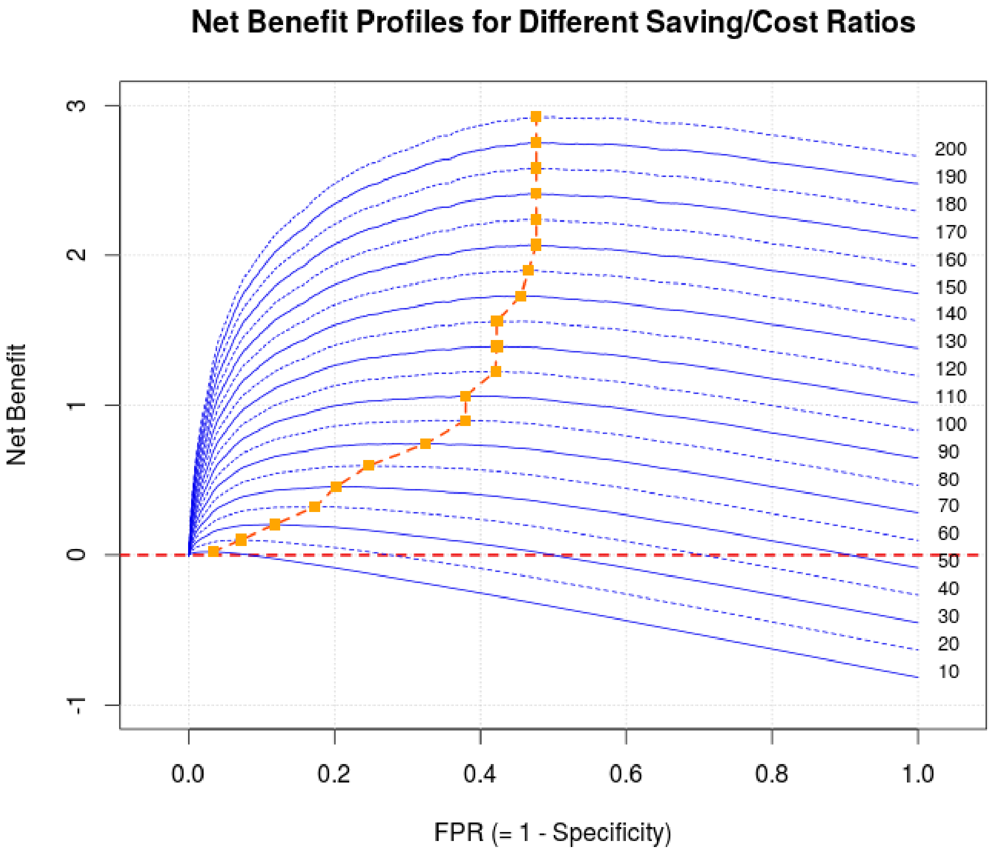
- For a threshold of 0.3 ==> Sensitivity = 0.72 (catching 244/1417)
- Threshold of 0.1 ==> Sensitivity = 0.522 (catching 739/1417)
- The second model would result in reaching out ~2500 more unnecessary customers by email, but also ~500 more of the customers who are really the goal of the effort (~ 3 times as much as the first case, and more than ½ of the total).

Model Evaluation - ROC

- The table shown in the previous slide can be summarized in a ROC Curve diagram, plotting together Sensitivity and Specificity as a function of Threshold (left panel below).
  - The dots and numbers labeling the colored lines mark cases for a few thresholds value.
- The right panel shows *Net Benefit Curves* for different ratios between Benefit (savings for avoiding a phone call processing) and Cost (of contacting the customers), for the same model, as a function of *False Positive Rate* (same variable of the X-axis of the ROC plot.)
  - Orange dots mark the highest point of each curve.



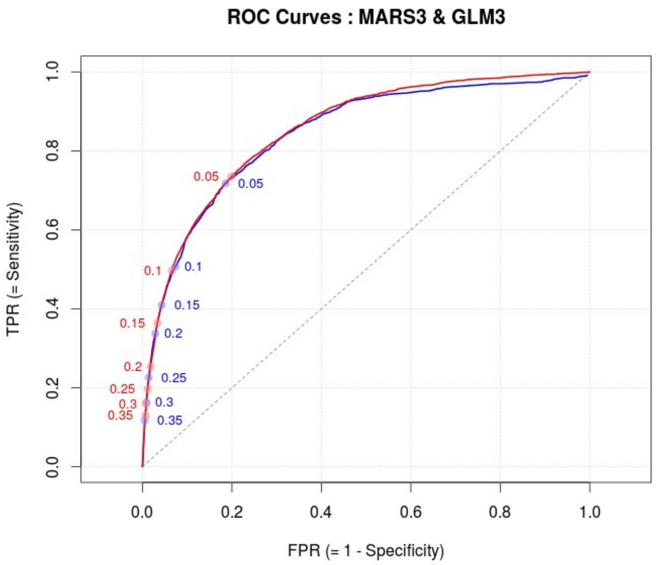
ROC Curves : MARS3 & GLM3



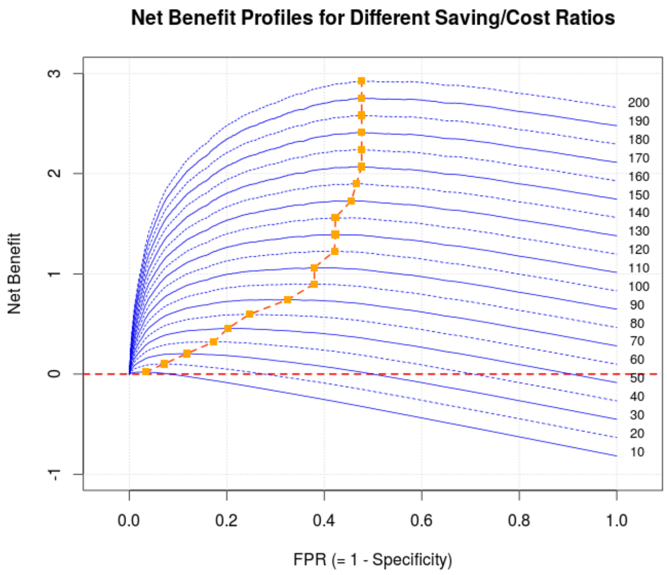
Net Benefit Profiles for Different Saving/Cost Ratios

Summary

- After evaluating each method on different sets of variables, on the basis of Sensitivity and Specificity, we deemed the MARS results to provide better predictive power.
- The inclusion of a few new variables summarizing (grouping) some of the numeric variables seems to add discriminatory ability to the model.
- At this stage *assuming* that the cost of sending more emails to customers who would not need to be encouraged to pay with self-service channels is significantly lower than the benefit (cost saving) of shifting a customer from service phone call to self-service, it seems possible to build a model on this dataset able to capture the majority of the customer “at risk” with a moderate increase of *false positives* (which represent a very low cost).
- For a more detailed, sophisticated and better supported recommendation it is necessary to take into consideration the various costs and benefits associated with the goal and the means to achieve it and evaluated different models on a “value matrix” basis.



ROC Curves : MARS3 & GLM3



Net Benefit Profiles for Different Saving/Cost Ratios