StepStone

ANALYSIS OF LEADS TO INCREASE THE EFFICIENCY OF THE SALES TEAM

By Alberto Escalante, Data Scientist (Applicant)

alberto.escalante@ini.rub.de

28.08.2018

CHALLENGE FACED BY THE SALES TEAM

- Sales Director is interested in increasing efficiency by understanding the leads
- 23,245 leads, € 18.9 M total contract value
- 13.5% of them result in a sale
- **Can we recognize the most promising leads to improve efficiency of the sales team?**
- **How likely is a lead to result in a sale?**
- **In case a lead results in a sale, how high will the sales value be?**
- How much revenue can be kept by reducing the leads to only 40%?

DATA AND METHODS

- 23,245 leads, 50 variables per lead
- Plus two key variables to be predicted:

1) Binary variable **Target_Sold** --> estimated using model 1 (M1)
2) Real valued variable **Target_Sales** (if contract) --> estimated using model 2 (M2)

- Use M1 and M2 to compute the *expected sales value* of each *lead*, defined as:

$$sales(\text{lead}) := \text{Pr}_{\text{M1}}(\text{Target sold} = 1 \mid \text{lead}) \times \text{Target sales}_{\text{M2}}(\text{lead})$$

FIRST MODEL (M1)

- Data --> Training set (60%), validation set (20%), test set (20%). Normalization
- Randomized search with 5-fold CV to tune hyperparameters
- Handling of missing values: replace by **zero** / mean. Optional: extend feature vector with Boolean flags that indicate whether a value is missing
- Four promising candidates

Algorithm	**Classification Rate (Validation)**
Logistic regression	90.88 %
Random forest classifier	**96.00 %**
Gradient boosting classifier	95.85 %
Support vector classifier	93.93 %
(Chance level)	86.28 %

SECOND MODEL (M2)

- Same data pre-processing / splitting
- Training data restricted to leads that resulted in a sale
- Five promising algorithms (randomized hyperparameter search + 5-fold CV)

Algorithm	Root Mean Square Error (Validation)
Linear regression	€ 11,205.7
Ridge regression	€ 11,215.2
Random forest regressor	**€ 8,652.3**
Gradient boosting regressor	€ 10,231.0
Support vector regression	€ 9,339.3
Chance level	€ 14,170.8

EXPECTED REVENUE (M1 & M2)

- Tested all combinations algorithms for m1 and m2
- Ordered the leads by expected revenue, kept the most promising 40% leads, computed total sales value for such leads
- Validation data € 3,506,349 in total

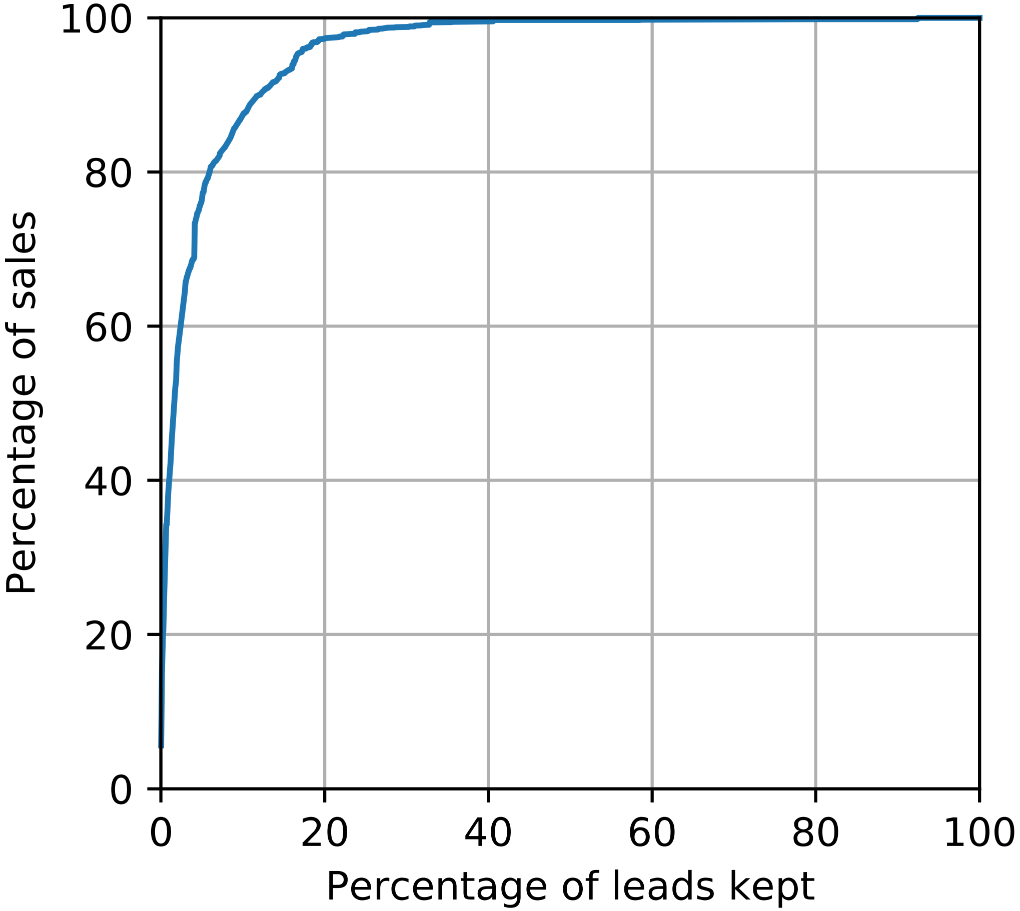
Percentage of sales kept (40% of leads)

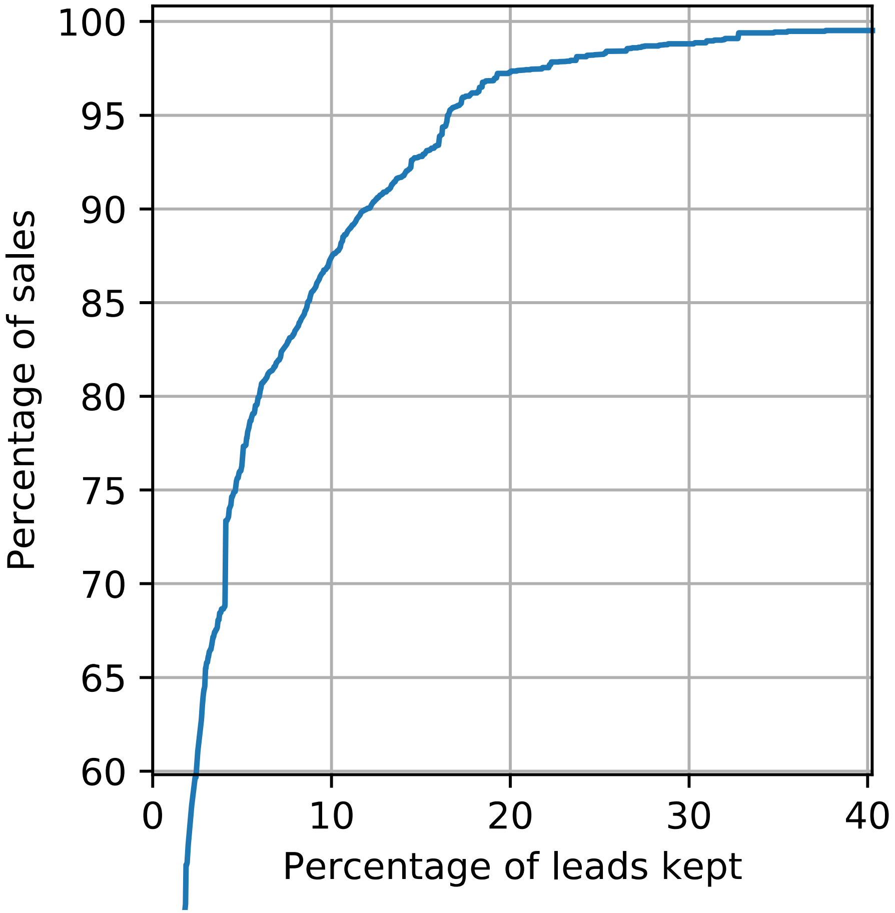
M1 **M2**	Logistic regression	Random forest classifier	Gradient boosting classifier	Support vector classifier
Linear regression	89.87 %	94.06 %	93.53 %	91.90 %
Ridge regression	90.79 %	93.31 %	93.83 %	92.13 %
Random forest regressor	98.48 %	**99.72 %**	98.90 %	97.79 %
Gradient boosting regressor	98.27 %	99.34 %	98.90 %	95.53 %
Support vector regression	96.71 %	99.13 %	98.71 %	96.57 %

EFFICIENCY GIVEN A FRACTION OF LEADS

StepStone

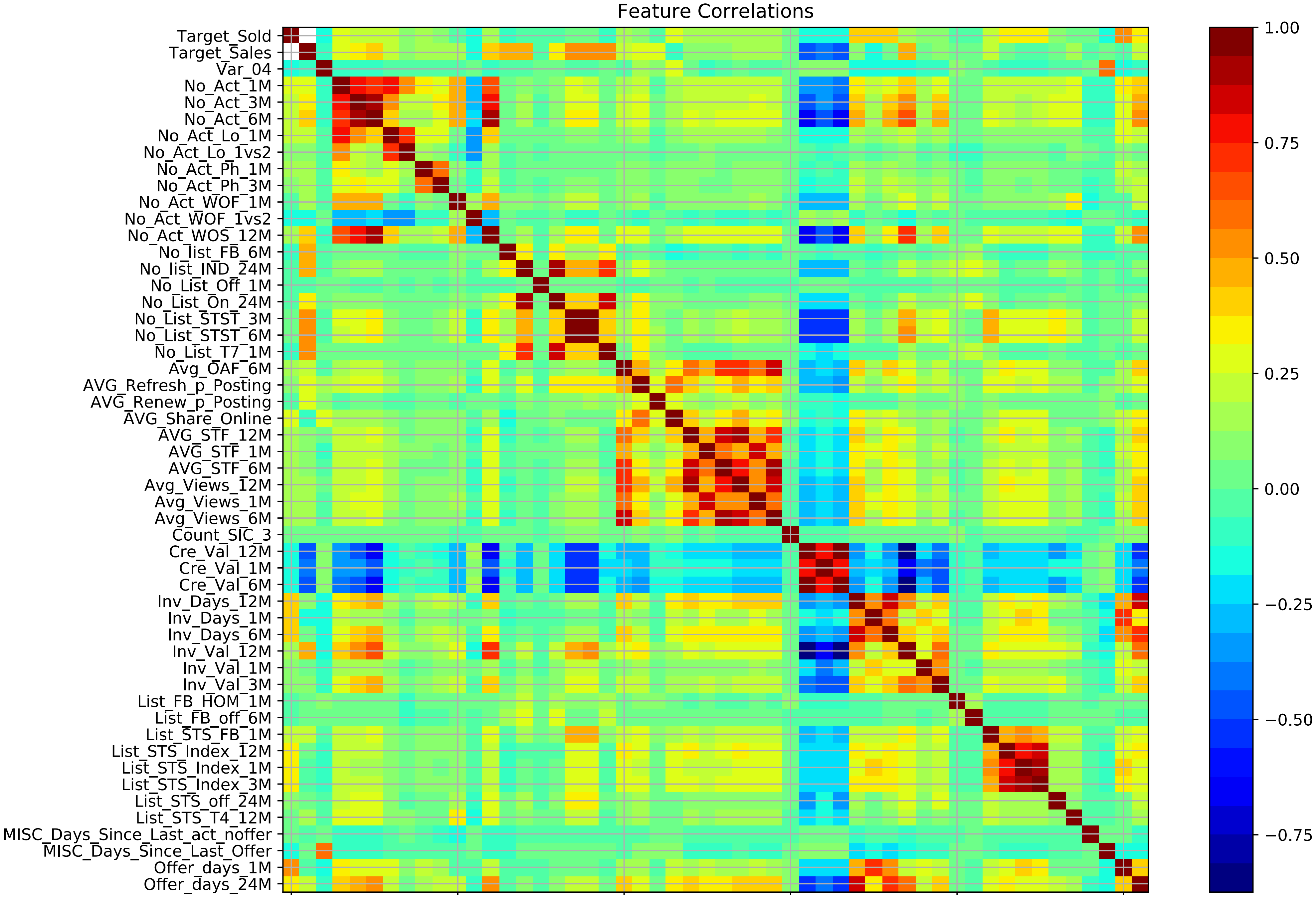
Best model: random forest classifier + random forest regression





VARIABLE IMPORTANCE (GINI, VAR) M1 & M2

Cre_Val_6M	**0.116**	Inv_Val_12M	0.018	No_List_On_24M	0.012	No_Act_WOF_1vs2	0.007
No_List_STST_6M	0.074	Inv_Days_1M	0.017	Inv_Val_1M	0.011	No_Act_6M	0.007
No_Act_Ph_1M	0.071	List_STS_Index_12M	0.017	AVG_Share_Online	0.011	List_STS_T4_12M	0.007
Cre_Val_12M	0.053	Inv_Val_3M	0.017	AVG_Refresh_p_Posting	0.011	No_Act_Lo_1vs2	0.006
MISC_Days_Since_Last_Offer	0.050	Offer_days_24M	0.016	Count_SIC_3	0.011	List_STS_off_24M	0.006
Cre_Val_1M	0.043	No_Act_3M	0.016	Avg_Views_1M	0.010	AVG_STF_1M	0.005
No_List_STST_3M	0.042	List_STS_Index_3M	0.016	Avg_Views_6M	0.010	List_FB_HOM_1M	0.004
Offer_days_1M	0.038	Inv_Days_12M	0.015	No_Act_WOS_12M	0.010	No_Act_WOF_1M	0.004
No_list_IND_24M	0.032	No_Act_1M	0.015	AVG_Renew_p_Posting	0.009	Var_04	0.004
Avg_OAF_6M	0.028	Inv_Days_6M	0.014	List_STS_FB_1M	0.009	List_FB_off_6M	0.002
Avg_Views_12M	0.027	No_Act_Ph_3M	0.014	AVG_STF_12M	0.008	No_List_Off_1M	0.001
MISC_Days_Since_Last_act_noffer	0.027	List_STS_Index_1M	0.013	No_List_T7_1M	0.007		
No_list_FB_6M	0.019	No_Act_Lo_1M	0.013	AVG_STF_6M	0.007		


Feature Correlations

StepStone

RECOMMENDATIONS AND FUTURE WORK

- Final evaluation using test data: € 4,338,430 (100% of leads) --> € 4,329,002 (99.78%) by pursuing only 40% of the leads
- Variable importance (average of M1, M2), (using gini/var metrics) is model dependent. Investigate whether important variables can be controlled to improve the expected sales value of a lead
- M1 and M2 give useful information to the sales team so that they can focus on the most promising leads
- Estimate the cost of pursuing a lead and then use the method to decide if it should be pursued
- Main limitation: the method is short sighted. Pursuing a lead (even if not immediate sold) might contribute to a conversion *in the future*. Care should be taken if the number of pursued leads is reduced
- Future work: variable selection, long-term effects, reinforcement learning