TΘΔ

Estimating customer satisfaction

C. SOFIA CARVALHO

NOKIA Software

R–Ladies Lisboa, 7 June 2018

Problem:

- Customer churning/attrition occurs when customers stop doing business with a company/service $\Rightarrow$ Loss of (revenue + customer loyalty).
- It is less expensive to retain existing customers than it is to acquire new customers $\Rightarrow$ Investment on retention of potential churners.
- Customers churn when they are dissastisfied with company/service.

Business interest:

Prevent customer churning by a) anticipating dissatisfaction and b) directing marketing effort to customer retention.

Goal:

Characterise and predict dissatisfied customers.

Plan:

1. Data preparation
2. Selection of independent variables
3. Classification methods
4. Sampling methods
5. Training + Testing
6. Performance metrics
7. Prediction of unclassified customers
8. Error in class prediction
9. Inferred function

1. Data preparation

DATA = Data set retrieved from network data base
+ Survey results retrieved from customer survey

⇒ `nrow` = 8226 customers, `ncol` = 978 variables.

- Remove variables with non–numerical or invalid data
- Replace missing values by the median of each variable
- Remove variables with zero variance

⇒ `ncol_val` = 824 valid variables.

- Separate customers with survey results (classified) from customers without survey results (unclassified)

⇒ `nrow_class` = 8213 classified customers,
`nrow_unclass` = 13 unclassified customers.

2. Selection of variables

Data set will most likely contain variables that are redundant or irrelevant to predict the survey results

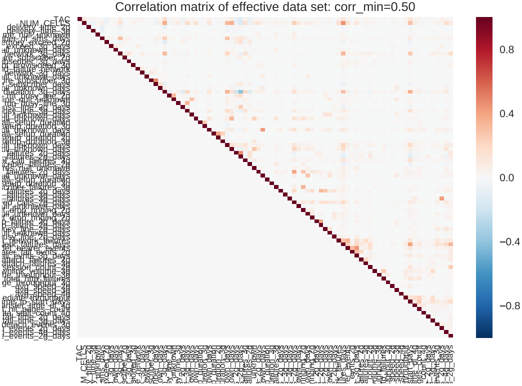
- Want to select the minimum set of variables that are **independent** and **relevant** to predict the survey results;
- Use either **a) the correlation between each pair of variables,** or **b) a classification method on the entire data set.**

NB: This selection criterion is agnostic with respect to the data–generating domain, hence applicable to data from any domain.

2.1 Correlation between pairs of variables

- To select the independent variables, set a minimum correlation value `corr_min` (e.g. 0.50) above which, from each pair of variables, keep one variable only;

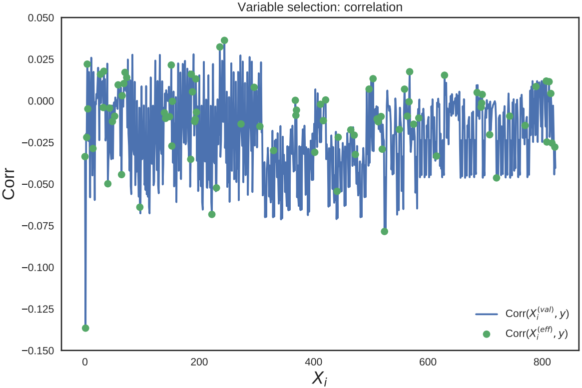
⇒ `ncol_eff` = 84 independent variables.



2.1 Correlation between pairs of variables

- To select the relevant variables, from each pair of dependent variables, keep the variable with the largest correlation with the survey results.

$\Rightarrow X_{ij}$: *i*=customer, *j*=variable



2.2 Logistic regression classifier

- Fit the data to the survey values using the logistic regression classifier $\Rightarrow$ outputs a score.
- To select the input parameters of the classifier, do a grid search of **a) the penalty parameter** $\mathtt{p} \in \{$**L1, L2**$\}$**,** which measures the sparcity, and of **b) the regularization parameter** $\mathtt{C}$, which measures the inverse of the regularization strength.
- Select the minimum set such that an increase of the number of variables yields a decrease in the score.
- Select the variables with non–vanishing coefficients as the relevant variables.

$\Rightarrow$ $\mathtt{ncol\_eff}$ = no. relevant variables.

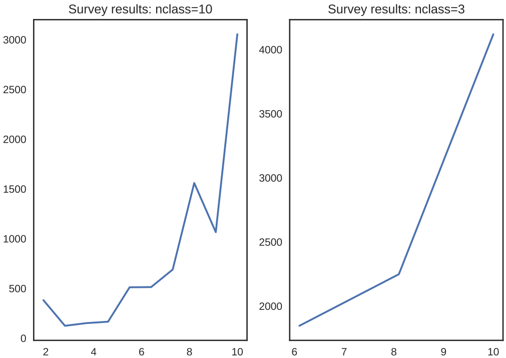
2.3 Re–bin survey values

Re–bin $\texttt{survey} \in \{1, 2, ..., 10\}$ into $\texttt{y} \in \{\texttt{ymin}, \texttt{ymed}, 10\}$:

- the dissatisfied customers: $\texttt{survey} \leq \texttt{ymin}$,
- the satisfied customers: $\texttt{ymin} < \texttt{survey} \leq \texttt{ymed}$,
- the very satisfied customers: $\texttt{survey} > \texttt{ymed}$.

Set $\texttt{ymin} = 6, \texttt{ymed} = 8$.

Fraction of customers over the three classes: $\{0.23, 0.27, 0.50\}$.



3. Classification methods

Assume that customers with similar patterns in the selected variables will also have similar values in the customer survey:

- Use a classification method to **infer patterns** in the selected variables;
- Use the inferred patterns to **predict the survey values.**

3.1 Classification methods: Examples

- K nearest neighbours
- Multi–layer perceptron
- **Random forest**
- Extra trees
- **Logistic regression**
- Stochastic gradient descent
- Support vector classifier

Criteria to select classifier: performance, computational cost, amenability to outputting analytical function.
Selected classifier: Logistic regression {`p`=L1, `C`=0.10}.

[Python: `http://scikit-learn.org/stable/modules/generated/`
R: `https://cloud.r-project.org/index.html`]

4. Sampling methods

- A typical data set on customer satisfaction is an example of an imbalanced data set, i.e. the class of satisfied customers outnumbers the class of dissatisfied customers.
- Most classification methods are built to:
  a) assume balanced classes and
  b) produce the simplest hypothesis that fits the data
  $\Rightarrow$ Imbalanced data sets are prone to a twofold bias in the classification results.
- **Possible solution:** generate synthetic data points belonging to the minority class (i.e. to over–sample the minority class) so as to balance the classes before applying a classification method.

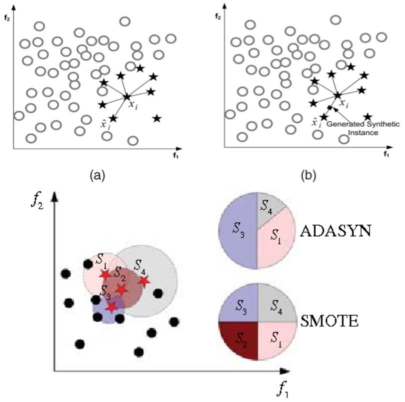
4.1 Sampling methods: SMOTE and ADASYN

- SMOTE: Synthetic Minority Over–sampling Technique

  [Chawla, Bowyer, Hall, Kegelmeyer (2002)]

- ADASYN: Adaptive synthetic sampling approach for imbalanced learning

  [He, Bai, Garcia, Li (2008)]



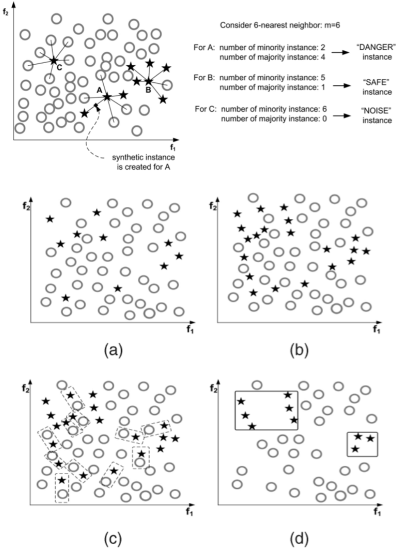
[Python: `http://contrib.scikit-learn.org/imbalanced-learn/stable/api.html`

R: `https://www.rdocumentation.org/packages/DMwR/versions/0.4.1/topics/SMOTE`]

4.2 Sampling methods: Variations of SMOTE

- Select points that act as synthetic sample generators according to ambiguity in point classification
  - borderline1 (smote_bl1);
  - borderline2 (smote_bl2);
  - svm (smote_svm);
- Combine minority class over–sampling with under–sampling
  - edited nearest neighbours (smote_enn);
  - Tomek links (smote_tomek);

[He, Garcia (2009)]



5.1 Training + Testing: Cross–Validation

- Cross–validation:
  - Divide the effective data set into $\mathtt{k}$ partitions, assigning: $\mathtt{k} - \mathtt{1}$ partitions $\rightarrow$ training set, 1 partition $\rightarrow$ test set.
  - Train the classification method on the training set, then test it on the test set by
    a) producing the probability of each survey value and
    b) predicting the survey value.
  - Rotate the assignment of the $\mathtt{k}$ partitions so as to
    a) train the method on different training sets and thus
    b) predict the survey value of all classified customers once.
- **NB1:** Divide the data so as to conserve in each partition the proportion among the classes observed in the entire data set.
- **NB2:** Over–sample the training set only: Involves comparing distances between points $\Rightarrow$ Must first normalize variables.

5.2 Training + Testing: Reshuffle cross–validation

- Repeat the cross–validation `sim` times, where each time the division of the effective data set results in different `k` partitions:

  `for isim in range(sim):`
    {Cross–validation
    `for ik in range(k):`
      {Over–sampling of training set}
      {Prediction of probability and survey value for:
        classified customers in test set : $P_{sc}(\hat{\mathtt{y}}_i = c | \mathtt{y}_i, \mathtt{x}_{ij}), \hat{\mathtt{y}}_i$
        unclassified customers: $P_{skc}(\hat{\mathtt{y}}_{i'} = c | \{\mathtt{y}_i, \mathtt{x}_{ij}\}, \mathtt{x}_{i'j}), \hat{\mathtt{y}}_{i'}$
      }
    }

  `sim` predictions for the classified customers,
  $\mathtt{sim} \times \mathtt{k}$ predictions for the unclassified customers.

6. Performance metrics

The prediction of the customer satisfaction is measured by performance metrics, namely:

- **Recall=TP/(TP+FN):** ratio of correct predictions of one class to the true size of the class → when do not want to wrongly classify as satisfied an dissatisfied customer and thus miss a potential churner;
- **Precision=TP/(TP+FP):** ratio of correct predictions of one class to the size of predictions of the class → when do not want to wrongly classify as dissatisfied a satisfied customer and thus use up churning resources with a satisfied customer.

The selection of performance metric depends on the business case, i.e. on balancing the cost of losing customers with the cost of targeting customers at risk of churning.

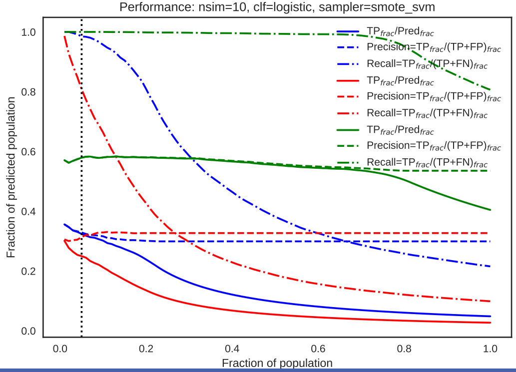
6.1 Performance metrics: Target prediction

Possible target prediction: correctly predict at least *perc* (e.g. 0.50) of *frac* (e.g. 5%) of the customers most likely to be in a given class:

$$\textbf{Performance metric}|_{frac} \equiv [\textbf{TP/(TP+TN+FP+FN)}]_{frac} \geq perc.$$

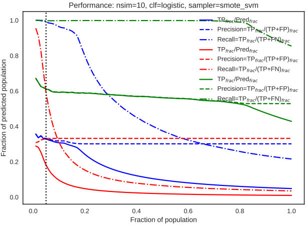
- Sort in decreasing order the probability of each customer belonging to each class.

`corr_min=0.50, ncol_eff=84:`

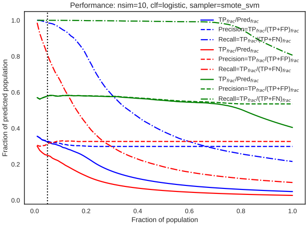


6.2 Best sampler

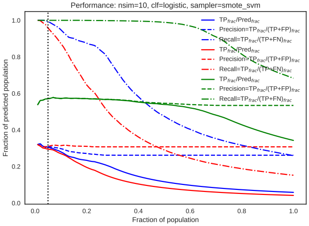
corr_min=0.30, ncol_eff=36



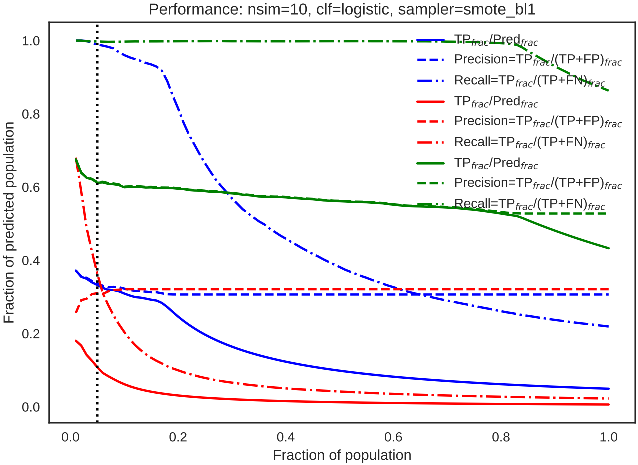
corr_min=0.50, ncol_eff=84



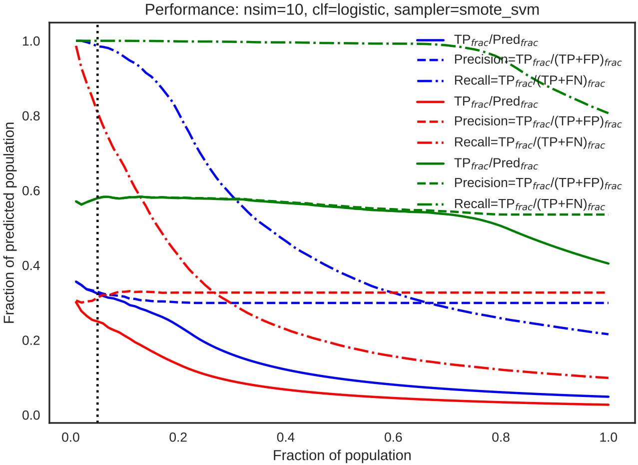
corr_min=0.70, ncol_eff=177



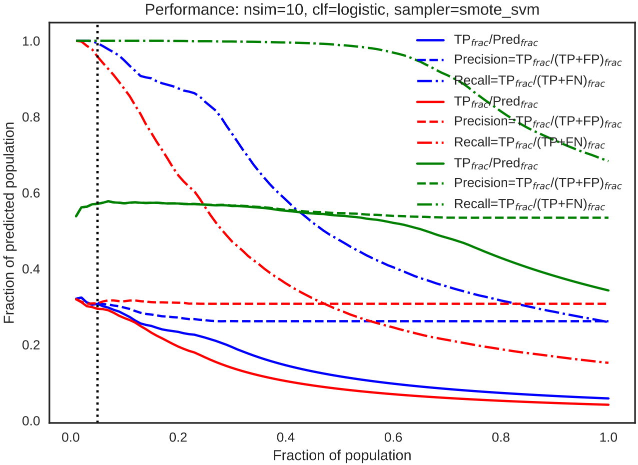
6.2 smote_bl1: corr_min=0.30, ncol_eff=36



6.2 smote_svm: corr_min=0.50, ncol_eff=84



6.2 smote_svm: corr_min=0.70, ncol_eff=177

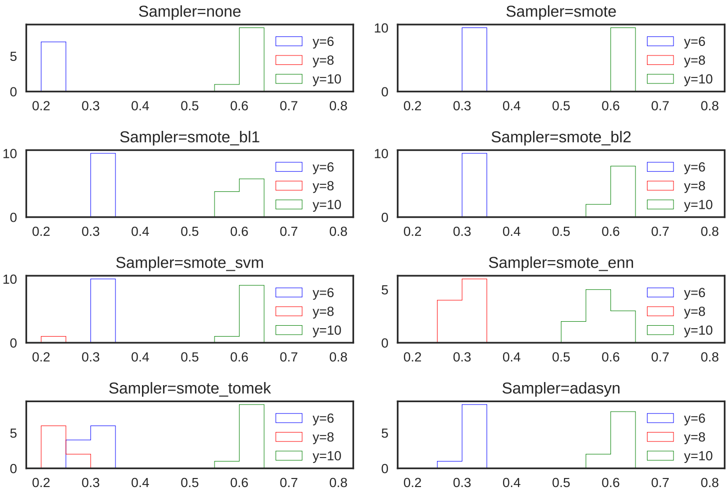


6.3 Performance metrics: Results

- Assuming interest lies in correctly identifying the dissatisfied customers, select the over–sampling method that yields the best performance at predicting $\mathtt{y} \leq \mathtt{ymin}$.
- Achieved performance $\sim$ 0.30 for the 5% customers most likely to be dissatisfied, which is better than random guess (0.23).

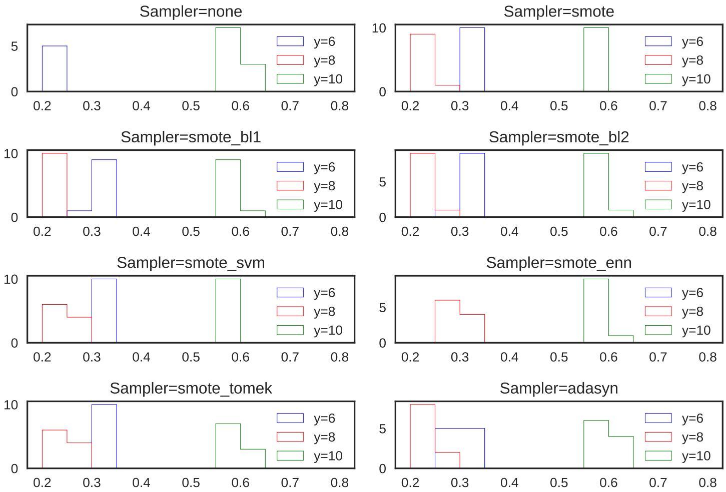
ncol_eff	Sampling method	$\sigma_{\mathtt{survey}}$	Target	Accuracy	$\mathtt{y} = 6$		$\mathtt{y} = 8$		$\mathtt{y} = 10$	
					Prec	Rec	Prec	Rec	Prec	Rec
36	smote_bl1	0	0.332	0.489	0.307	0.219	0.321	0.022	0.528	0.863
84	smote_svm	0	0.324	0.482	0.299	0.215	0.327	0.099	0.536	0.806
177	smote_svm	0	0.304	0.443	0.261	0.245	0.308	0.150	0.534	0.698

6.4 $perc = 0.5$: `corr_min=0.30, ncol_eff=36`



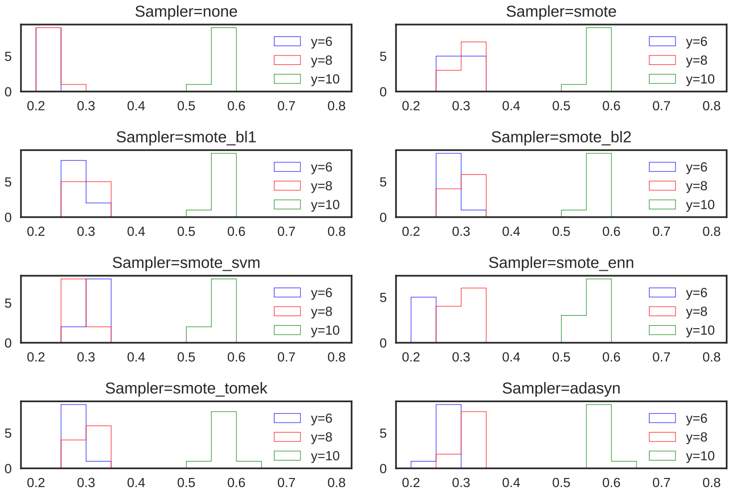
Best samplers: `smote_bl1`

6.4 $perc = 0.5$: `corr_min=0.50, ncol_eff=84`



Best sampler: `smote_svm`

6.4 $perc = 0.5$: `corr_min=0.70, ncol_eff=177`



Best samplers: `smote_svm`

7. Prediction of unclassified customers: $\mathrm{x}_{i'j}$, $i' \neq i$

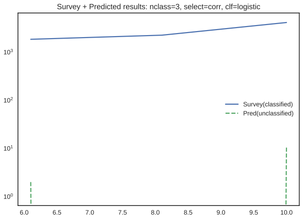
- For each s of the `sim` cross–validation runs, average over the `k` partitions (disjoint sets) with variance equal to the sample variance:

$$P_{sc}(\hat{\mathrm{y}}_{i'} = c|\{\mathrm{y}_i, \mathrm{x}_{ij}\}, \mathrm{x}_{i'j}) = \left\langle P_{skc}(\hat{\mathrm{y}}_{i'}|\{\mathrm{y}_i, \mathrm{x}_{ij}\}, \mathrm{x}_{i'j}) \right\rangle_k .$$

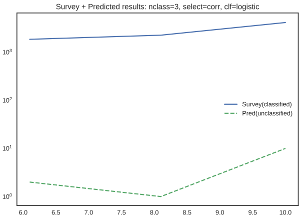
- Average over the `sim` runs (independent realizations of the training set) with variance equal to the inverted weighted sum:

$$P_c(\hat{\mathrm{y}}_{i'} = c|\{\mathrm{y}_i, \mathrm{x}_{ij}\}, \mathrm{x}_{i'j}) = \left\langle \left\langle P_{skc}(\hat{\mathrm{y}}_{i'}|\{\mathrm{y}_i, \mathrm{x}_{ij}\}, \mathrm{x}_{i'j}) \right\rangle_k \right\rangle_s .$$

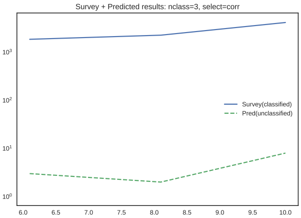
`corr_min=0.30, ncol_eff=36`



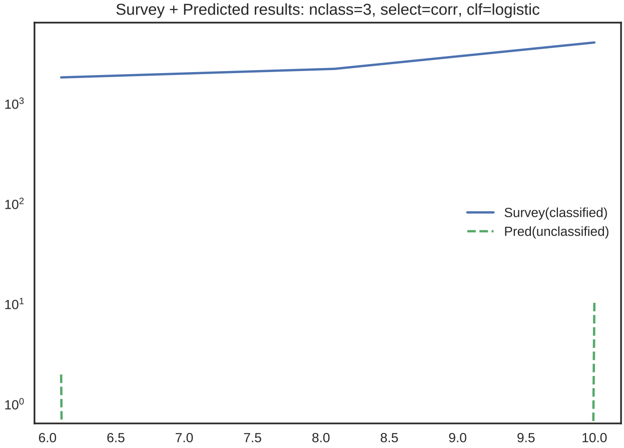
`corr_min=0.50, ncol_eff=84`



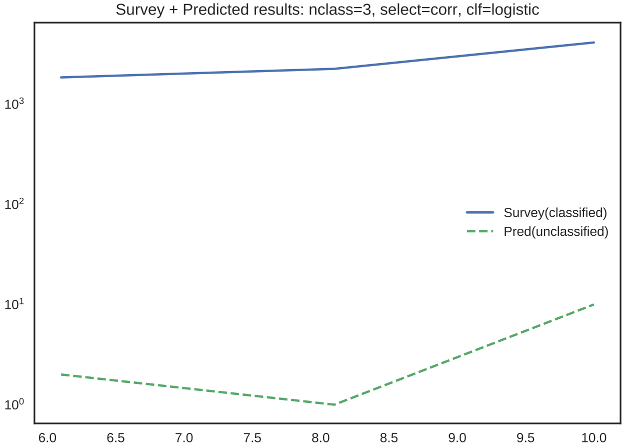
`corr_min=0.70, ncol_eff=177`



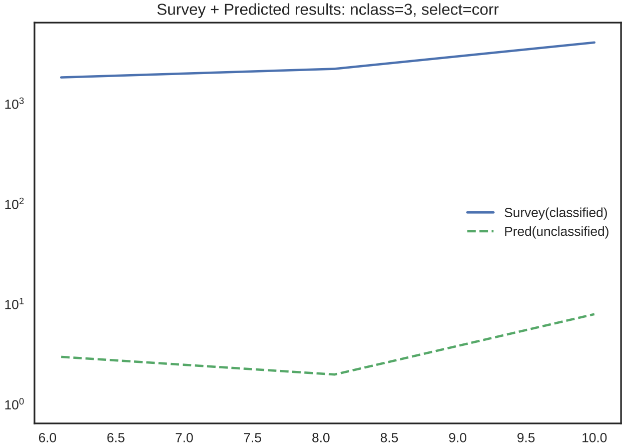
7.1 smote_svm: corr_min=0.30, ncol_eff=36



7.1 smote_svm: corr_min=0.50, ncol_eff=84



7.1 smote_svm: corr_min=0.70, ncol_eff=177



8. Error in class prediction

- The survey values have an intrinsic error given by $\sigma_{\texttt{survey}} = 2.47$.
- Recompute the performance metrics of the classification of $\texttt{X}_{ij}$ into three classes with respect to $\texttt{y}_{\texttt{pm}} =$ Re–bin$(\texttt{survey} \pm \sigma_{\texttt{survey}})$.
- Achieved performance ~ 0.60 for the 5% customers most likely to be dissatisfied, i.e. over twice as good as random guess (0.23).
- By considering $\texttt{y}_{\texttt{pm}}$ instead of $\texttt{y}$:
  $\Rightarrow$ TP and (TP+FN) increase, whereas (TP+FP) decreases
  $\Rightarrow$ recall stays practically unchanged, precision increases.

ncol_eff	Sampling method	$\sigma_{\texttt{survey}}$	Target	Accuracy	$\texttt{y} = 6$		$\texttt{y} = 8$		$\texttt{y} = 10$	
					Prec	Rec	Prec	Rec	Prec	Rec
36	smote_bl1	0	0.332	0.489	0.307	0.219	0.321	0.022	0.528	0.863
		2	0.651	0.761	0.623	0.201	0.456	0.017	0.792	1.
84	smote_svm	0	0.324	0.482	0.299	0.215	0.327	0.099	0.536	0.806
		2	0.629	0.748	0.613	0.199	0.440	0.072	0.795	1.
177	smote_svm	0	0.304	0.443	0.261	0.245	0.308	0.150	0.534	0.698
		2	0.599	0.709	0.557	0.236	0.443	0.117	0.794	1.

9.1 Inferred function: Classification output

At each s of the $\mathtt{sim}$ cross–validation runs, at each k of the $\mathtt{k}$ partitions, the outputs ($\mathtt{intercept}_{skc}, \mathtt{coef}_{skcj}$) of the classifier define an analytical function

$$f_{skc}(\mathtt{X}_{ij}) = \mathtt{intercept}_{skc} + \sum_{j=1}^{\mathtt{ncol\_eff}} \mathtt{coef}_{sckj}\mathtt{X}_{ij} \tag{1}$$

such that

$$P_{skc}(\hat{\mathtt{y}}_i = c|\mathtt{X}_{ij}) = \frac{1}{1 + \exp[-f_{skc}(\mathtt{X}_{ij})]} \tag{2}$$

is the probability that the predicted survey value $\hat{\mathtt{y}}_i$ of customer i is $\hat{\mathtt{y}}_i = c$.

9.2 Inferred function: Result

- Average the $\mathtt{sim} \times \mathtt{k}$ such functions f_{skc} to derive the resulting function

$$f_c(\mathrm{X}_{ij}) \equiv \left\langle \left\langle f_{skc}(\mathrm{X}_{ij}) \right\rangle_k \right\rangle_s \tag{3}$$

with variance

$$\sigma^2(f_{sc}) = \sigma^2_{\mathtt{intercept}_{sc}} + \sigma^2_{\mathtt{coef}_{sc}} \mathrm{X}^2,$$
$$\sigma^2_{f_c} = \sigma^2_{\mathtt{intercept}_c} + \sigma^2_{\mathtt{coef}_c} \mathrm{X}^2 + \sigma^2(f_{sc}).$$

- Use the resulting function to compute the probability that $\hat{\mathrm{y}}_{i'} = c$, given the values $\mathrm{X}_{i'j}$ for any new customers i', as

$$P_c(\hat{\mathrm{y}}_{i'} = c | \mathrm{X}_{i'j}) = \frac{1}{1 + \exp[-f_c(\mathrm{X}_{i'j})]} \tag{4}$$

with variance

$$\sigma^2_{P_c} = \left(\frac{\partial P_{\mathrm{c}}}{\partial f_c} \right)^2 \sigma^2_{f_c}.$$

Conclusions (α)

Concerning variable selection:

- Increase in the number of variables led to improvement in performance.
- $\mathtt{ncol\_eff} < 100$: class prediction is distributed between $\mathtt{y} = \mathtt{ymin}$ and $\mathtt{y} = \mathtt{10}$;
  $\mathtt{ncol\_eff} > 100$: class prediction is distributed across all three classes.

Concerning data over–sampling:

- Fixing the classification method and assuming a target prediction, over–sampling led to improvement of performance.

Conclusions (β)

Concerning performance measurement:

- Performance was evaluated:
  a) by assuming no error in the survey values and
  b) by assuming an error in the survey values or propagated through the inferred function.
- Case a): performance is very poor both in terms of precision and recall per class.
  Case b): performance improves in terms of precision per class.
- Dependence of the performance with the error is not only more statistically sound, but also an important piece of intelligence to be used in target negotiations with customers.

Conclusions (γ)

NB:

- Prediction is based on network data, hence encapsulates the "objective customer perception."
- Customer perception entangles an objective and a subjective perception.
- To estimate the "subjective customer perception," sentiment analysis might be relevant.