Predicting Customer Purchase to Improve Bank Marketing Effectiveness

Group 6

Sandy Wu, Andy Hsu, Wei-Zhu Chen, Samantha Chien

Business Goal

Problem

Re-calling “wrong” customers

- High labor costs
- Harming customer relationship

Business Goal

Improve marketing effectiveness by targeting the right customers

Stakeholders

- Bank Marketing Team
- Bank Service Employees
- Customers

Opportunities

- Gain revenues and lower costs by having more efficient marketing results

Challenges

- Using credit-scores
- Worsen the poor and rich disparity
- Very harmful for the mispredicted ones

Data Mining Goal

Data Mining Goal

Predict whether a certain customer will subscribe a term deposit or not

- Predictive, Forward-looking
- Supervised task
- Outcome variable : Subscribe/Not Subscribe
- Ranking (Find most likely subscribers)

Methods

Classification

- Naïve Bayes
- Logistic Regression
- Decision Tree
- Random Forest

Unbalanced Data

- SMOTE Oversampling

Performance

- ROC curves
- Lift Charts
- Sensitivity/Specificity
- F1-score

Data Description & Preparation

age	job	marital	education	default	housing	loan	contact	month	DOW	duration	campaign	pdays	previous	poutcome	emp.var.r	cons.price	cons.conf	euribor3m	nr.employ	y
60	technician	married	basic.4y	no	yes	no	cellular	oct	tue	595	2	6	1	success	-3.4	92.431	-26.9	0.737	5017.5	yes
70	retired	married	basic.4y	no	yes	no	cellular	oct	tue	76	2	999	0	nonexistent	-3.4	92.431	-26.9	0.737	5017.5	no
59	housemaid	married	basic.4y	no	yes	no	cellular	oct	tue	518	2	6	1	success	-3.4	92.431	-26.9	0.737	5017.5	yes
53	entrepreneu	married	basic.4y	no	yes	no	telephone	oct	tue	251	2	999	0	nonexistent	-3.4	92.431	-26.9	0.737	5017.5	no
30	technician	married	professiona	no	yes	no	cellular	oct	tue	427	1	999	0	nonexistent	-3.4	92.431	-26.9	0.737	5017.5	yes
24	managemer	single	university.c	no	yes	no	cellular	oct	tue	676	2	999	1	failure	-3.4	92.431	-26.9	0.737	5017.5	yes

Demographic data | Customer Credit data | Current Campaign data | Previous Campaign data | Social & Economic data

Data Source: UCI Machine Learning Repository

Data Size: 41,188 Rows, 21 Columns

Input Features: 'age', 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'DOW, 'campaign', 'pdays', 'previous', 'poutcome', 'emp.var.rate', 'cons.price.idx', 'cons.conf.idx', 'euribor3m', 'nr.employed'

Output Variable: y (subscribed: yes/no)

Partition: training/test = 0.7/0.3

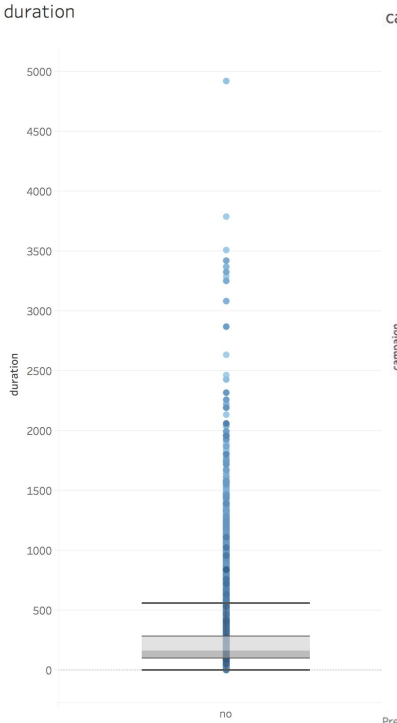
Data Prep:

1. normalization
2. dummies
3. pdays
4. duration

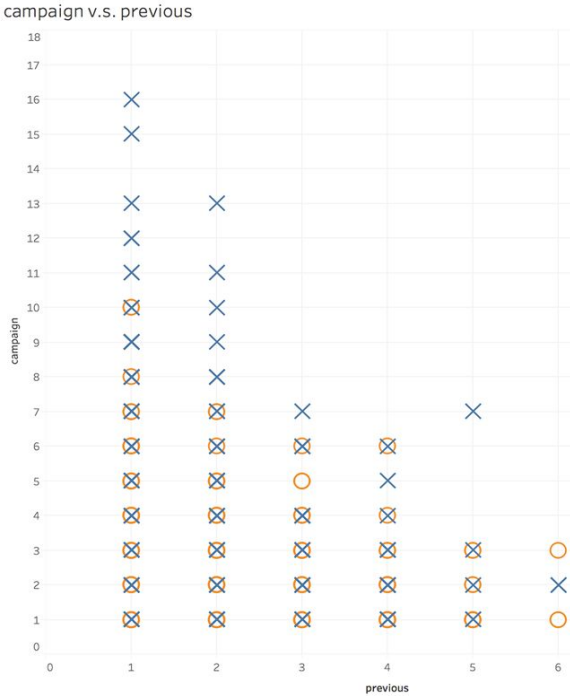
Training set SMOTE Oversampling (imbalance ratio = #0 / #1 = 790.27%)

Data Visualization

Duration / Output Bo

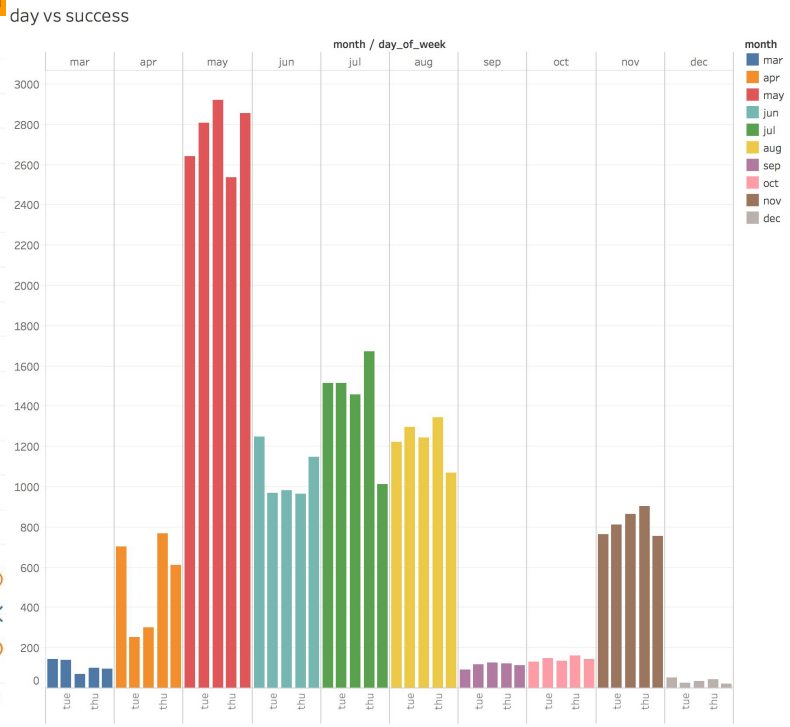


Previous / Campaign Scatter P



Previous vs. campaign. Color shows details about y. Size shows details about y. Shape shows details about y. The

DOW / Output Bar Chart



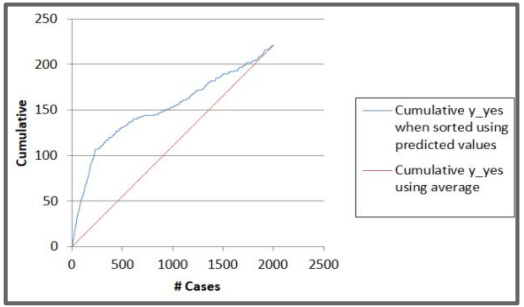
Distinct count of y for each day_of_week broken down by month. Color shows details about month.

Method Results

No Oversampled

Methods	Accuracy	Sensitivity	Specificity
Logistic Regression	64.47%	0.22	0.98
Decision Tree	66.04%	0.16	0.99
Naïve (Benchmark)	88.73%	0	1

Lift Chart of DT

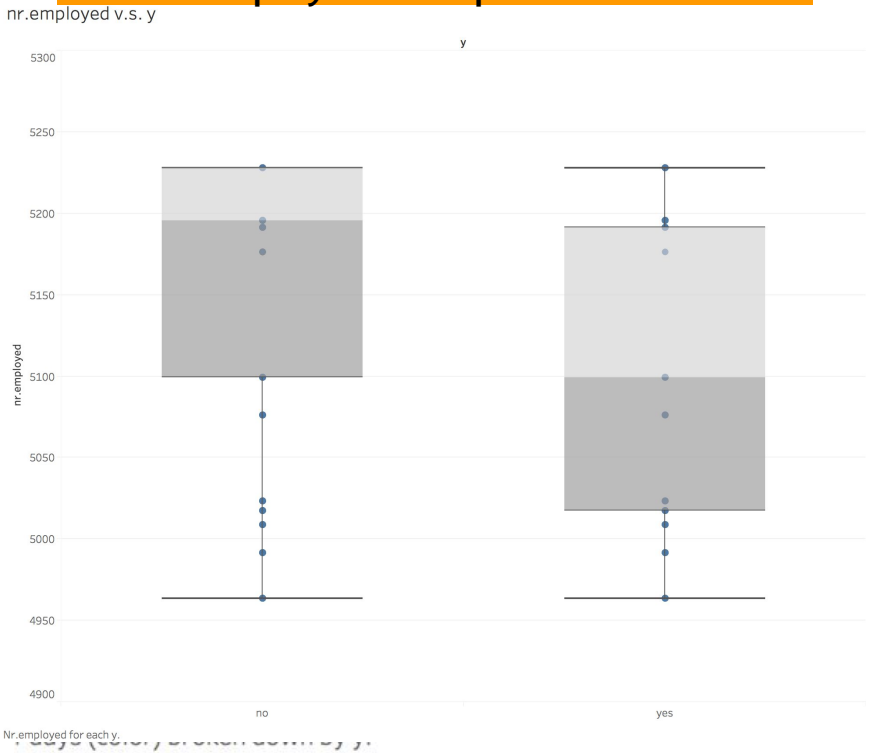


Oversampled

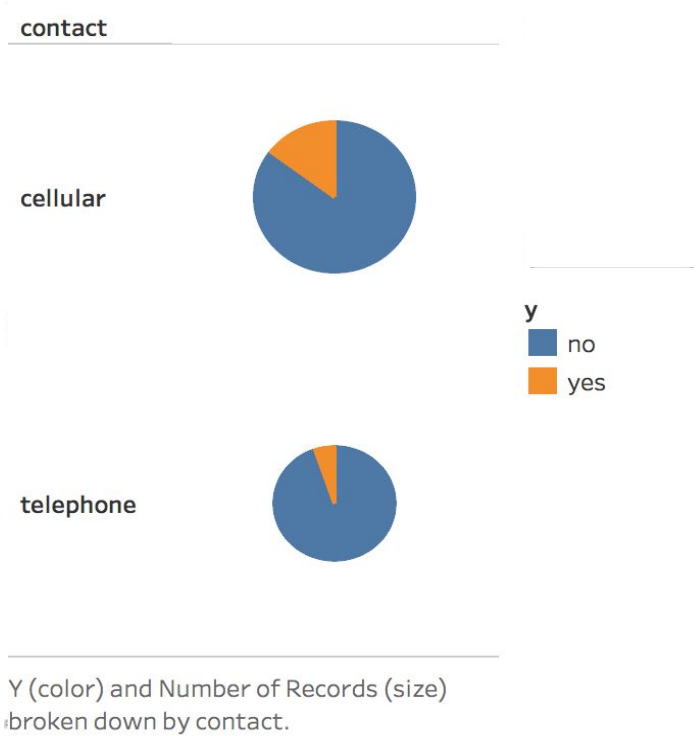
Methods	Accuracy	Sensitivity	Specificity	AUC	F1
Logistic Regression	81.57%	0.63	0.84	0.79	0.87
Decision Tree	85.60%	0.58	0.84	0.77	0.87
Random Forest	78.96%	0.64	0.81	0.79	0.87
Naïve Bayes	63.45%	0.75	0.62	0.76	0.80
Naïve (Benchmark)	88.73%	0	1		

Method Results (Oversampled)

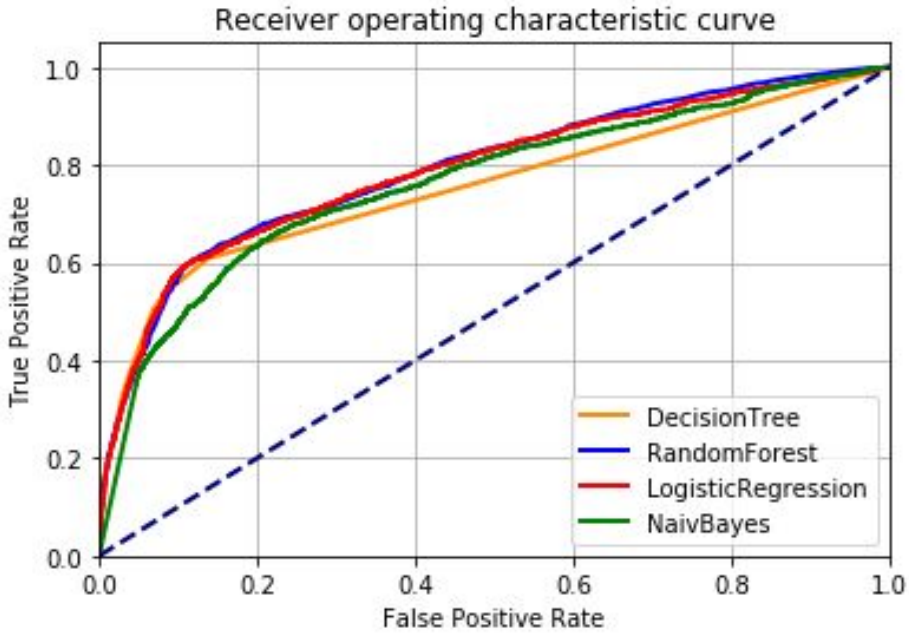
Nr.employed / Output Box Plot

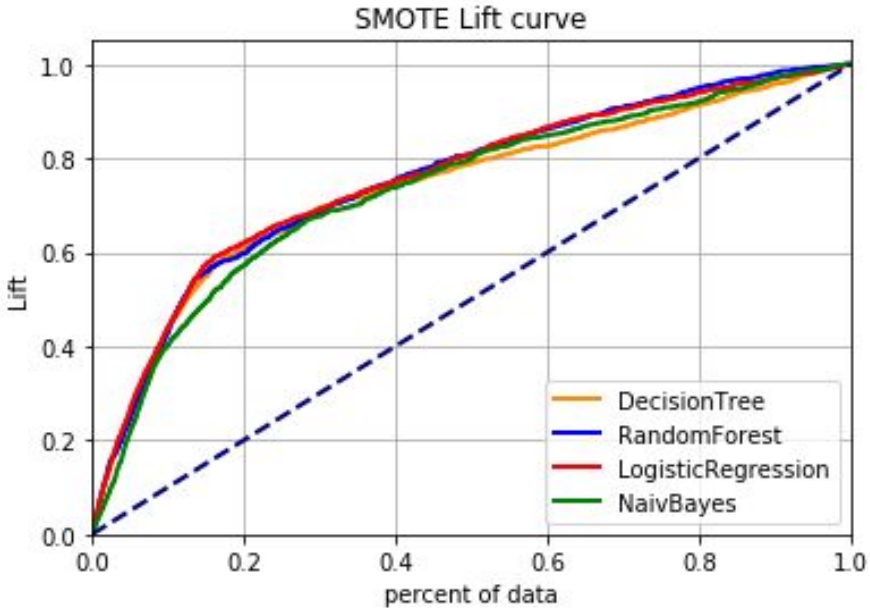


Contact / Output Pie Chart



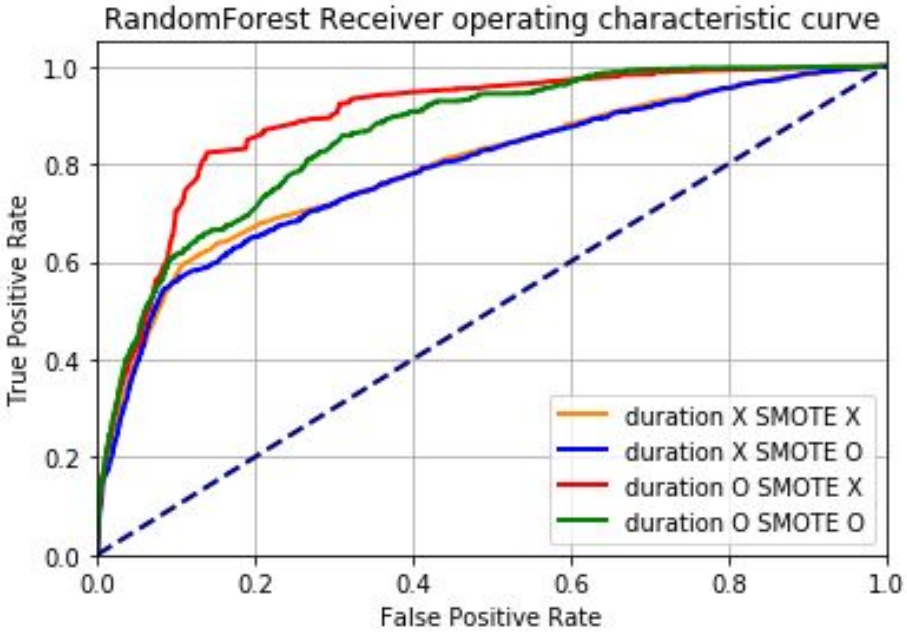
Performance Evaluation

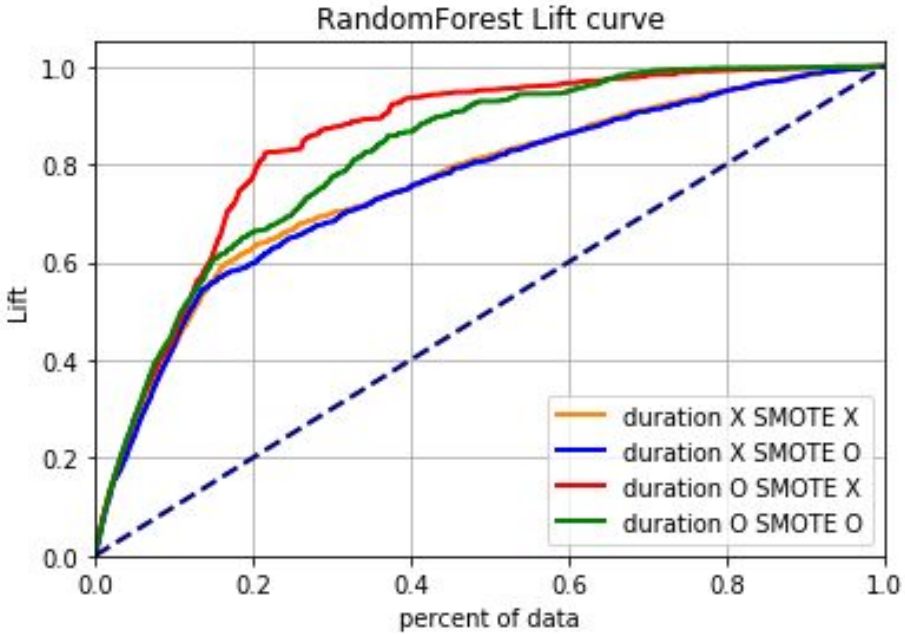




Other Findings & Comparisons

RandomForest in Different Conditions





Recommendations

- Features might have low correlations among them
  - Ask domain experts and include more related financial record columns
- More data samples may be better (~40,000 rows now)
- Including ordinal columns may bring about improvement in predictions