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

**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

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)

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

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 Box Plot

Previous / Campaign Scatter Plot
day vs success

DOW / Output Bar Chart
## Method Results

### No Oversampled

<table>
<thead>
<tr>
<th>Methods</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>64.47%</td>
<td>0.22</td>
<td>0.98</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>66.04%</td>
<td>0.16</td>
<td>0.99</td>
</tr>
<tr>
<td><strong>Naïve</strong> (Benchmark)</td>
<td>88.73%</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

### Lift Chart of DT

![Lift Chart](chart.png)

### Oversampled

<table>
<thead>
<tr>
<th>Methods</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>AUC</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>81.57%</td>
<td>0.63</td>
<td>0.84</td>
<td>0.79</td>
<td>0.87</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>85.60%</td>
<td>0.58</td>
<td>0.84</td>
<td>0.77</td>
<td>0.87</td>
</tr>
<tr>
<td>Random Forest</td>
<td>78.96%</td>
<td>0.64</td>
<td>0.81</td>
<td>0.79</td>
<td>0.87</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>63.45%</td>
<td>0.75</td>
<td>0.62</td>
<td>0.76</td>
<td>0.80</td>
</tr>
<tr>
<td><strong>Naïve</strong> (Benchmark)</td>
<td>88.73%</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Method Results (Oversampled)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.019621</td>
</tr>
<tr>
<td>Age</td>
<td>-0.146709</td>
</tr>
<tr>
<td>campaign</td>
<td>0.273018</td>
</tr>
<tr>
<td>pdays</td>
<td>-0.133803</td>
</tr>
<tr>
<td>previous</td>
<td>-3.050539</td>
</tr>
<tr>
<td>emp.var.rate</td>
<td>1.627592</td>
</tr>
<tr>
<td>cons.price.idx</td>
<td>0.179927</td>
</tr>
<tr>
<td>cons.conf.idx</td>
<td>0.460359</td>
</tr>
<tr>
<td>euribor3m</td>
<td>0.869257</td>
</tr>
<tr>
<td>nr.employed</td>
<td>-0.004862</td>
</tr>
</tbody>
</table>

Y (color) and Number of Records (size) broken down by contact.
Performance Evaluation
Other Findings & Comparisons

RandomForest in Different Conditions

![RandomForest ROC Curve](image1)

![RandomForest Lift Curve](image2)
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