Identifying the Factors Responsible for Loan Defaults and Classification of Customers using SAS® Enterprise Miner

Juhi Bhargava | Prashanth Reddy Musuku
Introduction
## Data Collection and Preparation

- Lending Club Loan dataset for 2015

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Final</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observations</td>
<td>91,233</td>
</tr>
<tr>
<td>Variables</td>
<td>96</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Excluded variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Records with same value</td>
</tr>
<tr>
<td>Similar variables like ID, Member_ID</td>
</tr>
<tr>
<td>Too many levels</td>
</tr>
<tr>
<td>Percentage of missing values &gt; 50%</td>
</tr>
<tr>
<td>Variable with supplementary information</td>
</tr>
</tbody>
</table>
Data Exploration

Loan Status

- Default: 0.02%
- Late (16-30 days): 2.60%
- In Grace Period: 4.66%
- Late (31-120 days): 9.61%
- Charged Off: 17.34%
- Fully paid: 65.77%
Data Exploration

Distribution of Loan Purpose

- Debt consolidation: 64.34%
- Credit card: 19.96%
- Home improvement: 6.66%
- Small business: 1.12%
- Car: 0.57%

Total Funded Amount: $1,364,126,925.00
Data Exploration

Interest Rate vs Average Total Received Principal

- 5-10%:
  - 36 months: $15,880
  - 60 months: $13,950

- 10-15%:
  - 36 months: $8,156
  - 60 months: $11,414

- 15-20%:
  - 36 months: $6,211
  - 60 months: $10,838

- 20-25%:
  - 36 months: $4,370
  - 60 months: $10,184

Term:
- 36 months
- 60 months
Data Exploration

Grade vs Average Loan Amount

<table>
<thead>
<tr>
<th>Grade</th>
<th>36 months</th>
<th>60 months</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$13,724</td>
<td>$18,621</td>
</tr>
<tr>
<td>B</td>
<td>$12,154</td>
<td>$20,093</td>
</tr>
<tr>
<td>C</td>
<td>$11,716</td>
<td>$19,953</td>
</tr>
<tr>
<td>D</td>
<td>$12,626</td>
<td>$19,774</td>
</tr>
<tr>
<td>E</td>
<td>$13,563</td>
<td>$20,709</td>
</tr>
<tr>
<td>F</td>
<td>$11,272</td>
<td>$22,038</td>
</tr>
<tr>
<td>G</td>
<td>$10,037</td>
<td>$22,154</td>
</tr>
</tbody>
</table>

Term
- 36 months
- 60 months
Data Preparation

- Loan Status converted into Good loans and Bad loans

Diagram:
- Loan Status
  - Fully Paid
  - In Grace Period
  - Late (16-30) Days
  - Late (31-120) Days
  - Default
  - Charged Off

- Good Loan (73%)
- Bad Loan (27%)
Data Preparation, Variable Clustering and Selection

**Data Preparation**
- Imputation on variables - last payment date, %BC accounts greater than 75%.
- Transformation carried out on variables
- Data Partition - 70% training and 30% validation

**Variables Selected**
- Collection Recovery Fee
- Last Credit Pull Date
- Last Payment Date
- Total Received Principal
- Outstanding Principal Invested
- Total Bank card Credit Limit
- Term
- Verification Status
- Grade
- Total High Credit Limit
- Percent of Bankcard Accounts > 75% of Limit
Modeling

Logistic Regression

<table>
<thead>
<tr>
<th>Influential Variables</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>6.9077</td>
</tr>
<tr>
<td>Total Principal Received</td>
<td>-17.1923</td>
</tr>
<tr>
<td>Collection Recovery Fee</td>
<td>66.7564</td>
</tr>
<tr>
<td>Outstanding Principal Remaining</td>
<td>2.8337</td>
</tr>
<tr>
<td>Last Payment Date (July 16)</td>
<td>-1.8584</td>
</tr>
</tbody>
</table>

Neural Networks

- **Important Variables**
  - Total Principal Received
  - Outstanding Principal
  - Last Payment Date

| Misclassification Rate                   | 0.07672          |

Random Forest

- **Important Variables**
  - Total Principal Received
  - Total BC Limit
  - Verification Status
  - Last Payment Date

| Misclassification Rate                   | 0.05619          |

Decision Tree

- **Important Variables**
  - Total principal received to date
  - Last payment date
  - Remaining outstanding principal
  - Loan credit pull for last payment
  - Term

| Misclassification Rate                   | 0.05966          |
Model Evaluation

<table>
<thead>
<tr>
<th>Model</th>
<th>Misclassification Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>0.05619</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.05966</td>
</tr>
<tr>
<td>Neural Network</td>
<td>0.07672</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.08016</td>
</tr>
</tbody>
</table>

**Future Approach**

- Analysis based on Customers’ demographics
- Comparison of loans based on region
Juhi Bhargava
juhi.bhargava@okstate.edu
(405) 780 5640
www.linkedin.com/in/juhibhargava

Prashanth Reddy Musuku
prashanth.musuku@okstate.edu
(732) 770 3399
www.linkedin.com/in/prashanthreddymusuku