Implementing a Scorecard Model for Financial Data

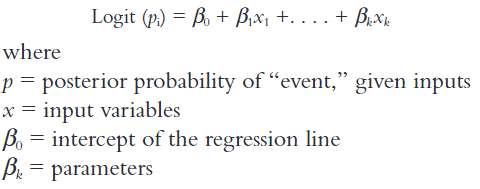
1. **Data Analysis**
   1. Missing Values/Outliers - Must deal with to do regression
      1. Several options for missing values – best to categorize missing values as separate attribute
      2. Other options:
         1. Could exclude features with missing values – may remove a lot of data
         2. Could impute missing values
   2. Correlation
      1. **Example**: “PROC VARCLUS” on SAS “uses a type of principal components analysis to identify groups of characteristics that are correlated. One can then select one or more characteristics from each group, and theoretically, represent all the information contained in the other characteristics in each of the groups.” [Siddiqi 77]
         1. Better than just using correlation – since it considers collinearity as well as correlation
   3. [Multicollinearity is bad, but you don’t necessarily need to remove it](https://stats.stackexchange.com/a/268975)
      1. Large sample sizes mitigate the effects of MC (Multicollinearity)
2. **Statistical Measures**
   1. Definitions
      1. Characteristic – one value/variable/column of the data set
      2. Attribute – one WOE bin of a characteristic
   2. **WOE** - Generate WOE bins and values for **continuous** characteristics
      1. Distribution of Goods - % of Good Customers in a particular group
      2. Distribution of Bads - % of Bad Customers in a particular group
      3. WOE = In(% of non-events ➗ % of events)
      4. Separate bin for “Missing” groups
      5. General rule: “minimum 5% of data in each bucket”
      6. No groups without either good or bad values (divide by zero)
      7. Each bin has sufficiently different bad rate and WOE
      8. Bin WOE pattern should usually be monotonic (Or have **logical** relationships)
         1. Exceptions to be judged by person with business intelligence (exceptions: Siddiqi 84)
      9. **Nominal** (categorical/non-continuous) characteristics are grouped by combining attributes with similar WOE
   3. **IV** - Generate IV (Information Value) for characteristics
      1. IV = ∑ (% of non-events - % of events) \* WOE
      2. Indicates predictive power of characteristic for the data
      3. Usually use Medium/Strong variables

|  |  |
| --- | --- |
| **Information Value** | **Variable Predictiveness** |
| Less than 0.02 | Not useful for prediction |
| 0.02 to 0.1 | Weak predictive power |
| 0.1 to 0.3 | Medium predictive power |
| 0.3 to 0.5 | Strong predictive power |
| Greater than 0.5 | Suspicious predictive power |

1. **Preliminary Overview**
   1. General rule: between 8 and 15 characteristics
      1. Characteristics represent as many independent types of data as possible
   2. Review of scorecard concept
      1. “Mimic the thought process of a seasoned, effective adjudicator or risk analyst”
      2. Forms a “risk profile” of the subject
   3. Subsequent monitoring of the scorecard is a priority
      1. As the population (demographics, behavior, culture) changes, the scorecard’s accuracy will change
      2. Monthly reports to monitor scorecard accuracy
2. **Logistic** **Regression** - Perform logistic regression using WOE values
   1. *Linear* regression is used for a continuous-variable target
      1. Linear regression is not capable of predicting probability
   2. *Logistic* regression is used when the target variable is categorical
      1. Models the *probability* that a given output belongs to a given category
      2. Restricts predicted output to between 0 and 1
   3. Logistic function derivations (using only one input variable, X)

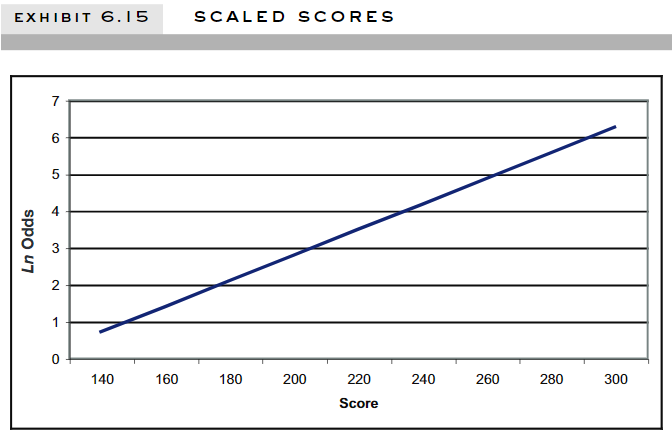
|  |  |
| --- | --- |
| Logistic Function | Reframed Logistic Function |
|  |  |
| **Odds Ratio** defined as | Logit defined as |
|  |  |

* 1. Logit when using multiple variables



* 1. Use WOE of each grouping as the input
  2. Finding the best possible model (best combination of variables) using regression
     1. “All possible” regression – exhaustive search – usually too computationally expensive
     2. Three types of stepwise logistic regression techniques
        1. Forward Selection
           1. Add best characteristics (by predictive power e.g. p-value, Chi Square, etc.) one by one to model until a certain threshold.
           2. i.e. Start with the #1 characteristic in the model. Then test adding each other characteristic individually/one-by-one to the model. Select the characteristic that causes the model to perform best as the second variable.
           3. Continue this process to select more characteristics until a desired performance measure is reached.
        2. Backwards Elimination
           1. Opposite of forward selection
           2. Start with all characteristics in the model. Remove characteristic one-by-one based on predictive power
        3. Stepwise
           1. A combination of forward selection and backwards elimination
           2. Alternate steps each of the two techniques
  3. Using regression to evaluate scorecards
     1. Model-ordering option in stepwise regression: two choices
        1. Single regression
        2. Multiple regression

1. **Reject Inference**
   1. So far, we only have data from loans that were approved and given
   2. It is also important to account for cases of people who have been rejected
      1. May be difficult to consolidate such information
2. **Designing the Scorecard [Siddiqi 92-119]**
   1. Scaling scores
      1. Scores can be in many different forms
         1. Odds ratio
         2. “Defined min/max scale with a specified odds ratio at a certain point and specified rate of change of odds [Siddiqi 114]
      2. Scaling is purely cosmetic – does not affect performance
         1. Consider implementability with other tools, ease of understanding, and continuity with other scorecards
      3. A common scale used in industry
         1. “A scorecard with discrete scores scaled logarithmically, with the odds doubling at every 20 points”

****

* + 1. Scaling calculation
       1. Score = Offset + Factor\*ln(odds)

1. **Model evaluation**
   1. Cross-validation
   2. Confusion matrix
   3. ROC (Receiver Operating Characteristic) & AUC (Area Under Curve)
      1. ROC shows the classification power of the model compared to a purely random classifier
      2. The AUC is the integral of the ROC, numerically representing the models’ classification power
   4. Reject inference
      1. We do not have binary data on rejected applicants
      2. Seeks to augment data from rejected applicants