Predicting Loan Defaults using Predictive Analytics

DAT 690- Data Analytics Capstone

SNHU

By LIKANE ANNE DRUIDE

10/12/2019

Introduction

Background:

- **Company**: GE credit department provides credit and financing solutions.
- **Current situation**: GE is approving credit requests based on the applicants’ socio demographic and financial data.
- **Challenge**: How to limit the credit risk associated with the loans.
- **Solution** : Use Predictive analytics to predict future loan defaults.

Introduction- Cont’d

Significance of the predictive model in solving the problem:

-Identify the variables that have the most effect on people’s ability to pay back their debts.

-Help define a rigorous approach to how credit/loan is approved.

-Improve risk management.

Pilot Evaluation

- A Predictive model was built using data from 1,000 past customers.

- Success: Model can determine which applicant will default with an 80.7% accuracy level.

Current practice	With predictive capabilities
Loan/Financing decisions based on financial, socio demographics of applicants.	Loan/Financing decisions based on financial, socio demographics of applicants with known (estimated) risk level.
NA	Prediction of future loan defaults
NA	Model Results are 80.7% accurate

- Challenge: Identify the most appropriate predictors of credit risks. Data privacy and security

Predictive Modeling : Cost vs Benefits

Cost

- Low implementation costs as level of automation is high.
- Free software (open source - statistical software).

ROI

- Use of model will provide a competitive advantage to GE (insights into future events and can approve more applicants).
- Use of model will allow to better assess and manage credit risk (financial loss reduced).

Plan Modification

Addresses the concerns from pilot : data security, predictors identification

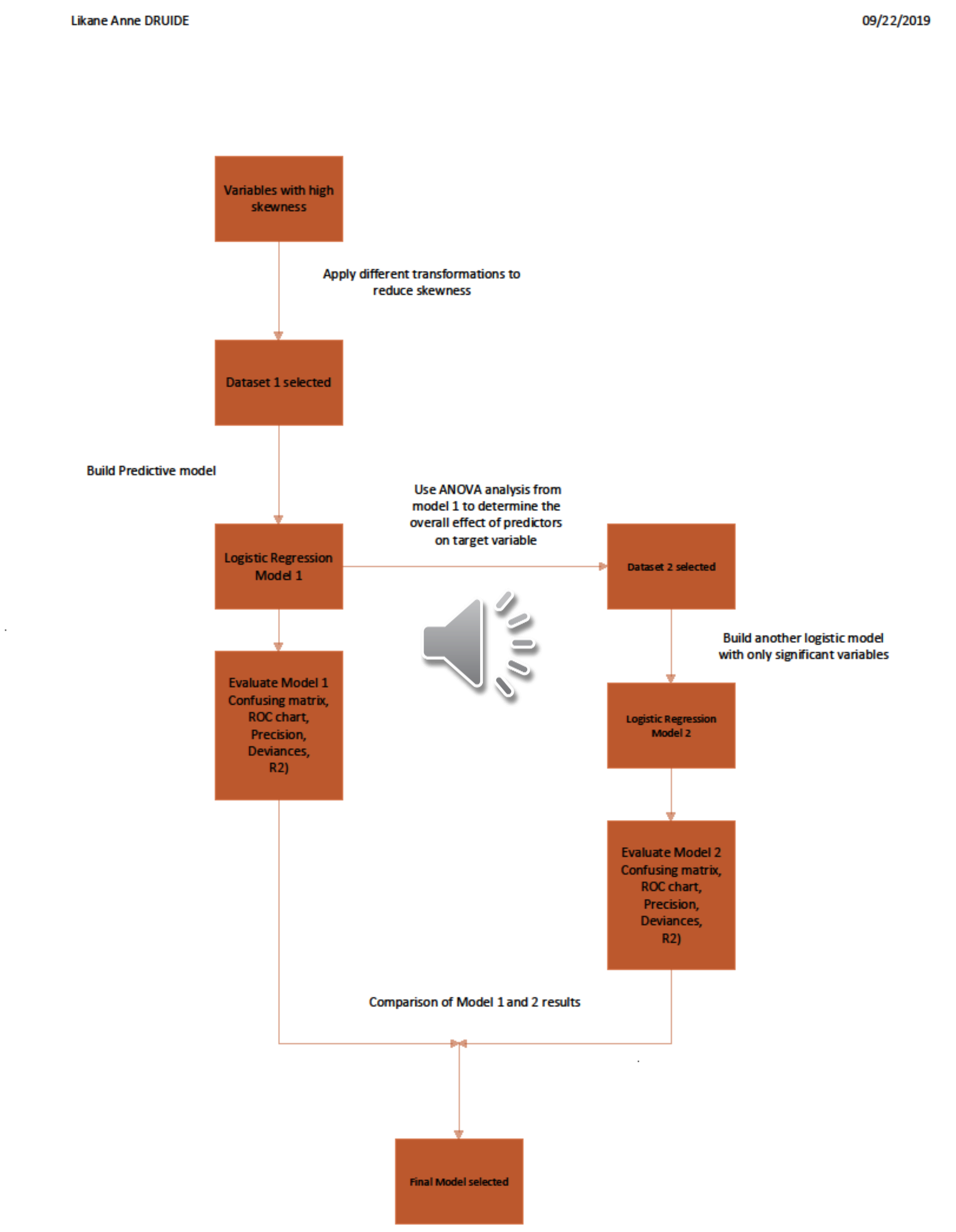
Data Security and privacy

- Informed consent prior to data collection.
- Regulatory compliance requirements for financial institutions.
- Access control
- Data encryption
- Data stored on secure servers

Variables selection

- Data quality: check for inconsistency, incompleteness, accuracy and missing/unknown data points.
- Improve data understanding

Modified Analytic plan



Likane Anne DRUIDE

09/22/2019

Variables with high skewness

Apply different transformations to reduce skewness

Dataset 1 selected

Build Predictive model

Logistic Regression Model 1

Use ANOVA analysis from model 1 to determine the overall effect of predictors on target variable

Dataset 2 selected

Build another logistic model with only significant variables

Logistic Regression Model 2

Evaluate Model 1 Confusing matrix, ROC chart, Precision, Deviances, R2)

Evaluate Model 2 Confusing matrix, ROC chart, Precision, Deviances, R2)

Comparison of Model 1 and 2 results

Final Model selected

Model Implementation : Logistic Regression

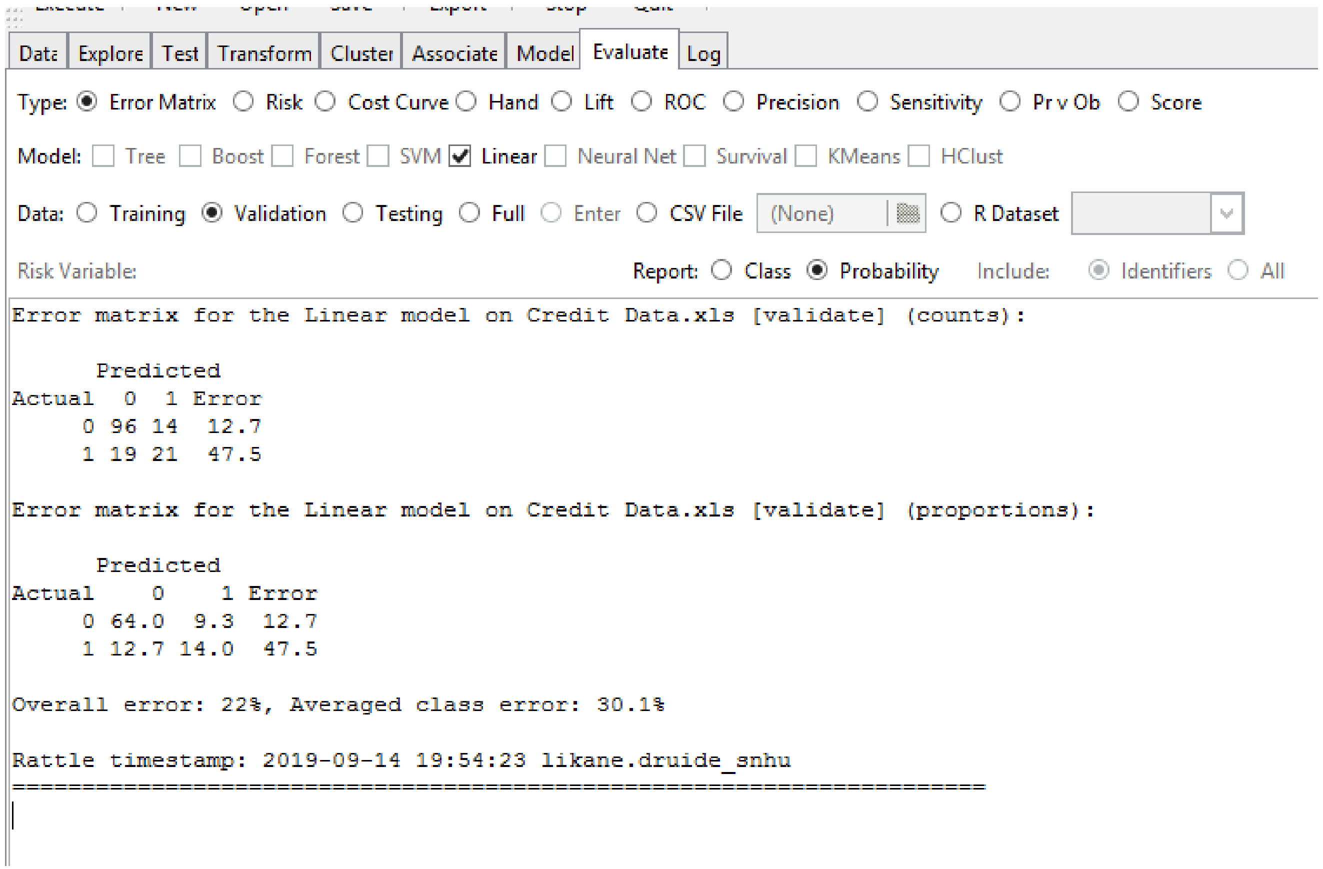
Analytic tool

- R and Rattle. R is free powerful statistical software. Rattle is an easy to use *Graphical* User Interface (*GUI*) for the R software.
- These tools allow to analyze large datasets to create data visualizations and to build models.

Logistic Regression Model Results

- Confusion Matrix

Model can predict that an applicant will default on his payment with a 78% accuracy.



Type: ● Error Matrix ○ Risk ○ Cost Curve ○ Hand ○ Lift ○ ROC ○ Precision ○ Sensitivity ○ Pr v Ob ○ Score

Model: □ Tree □ Boost □ Forest □ SVM ☑ Linear □ Neural Net □ Survival □ KMeans □ HClust

Data: ○ Training ● Validation ○ Testing ○ Full ○ Enter ○ CSV File (None) ○ R Dataset

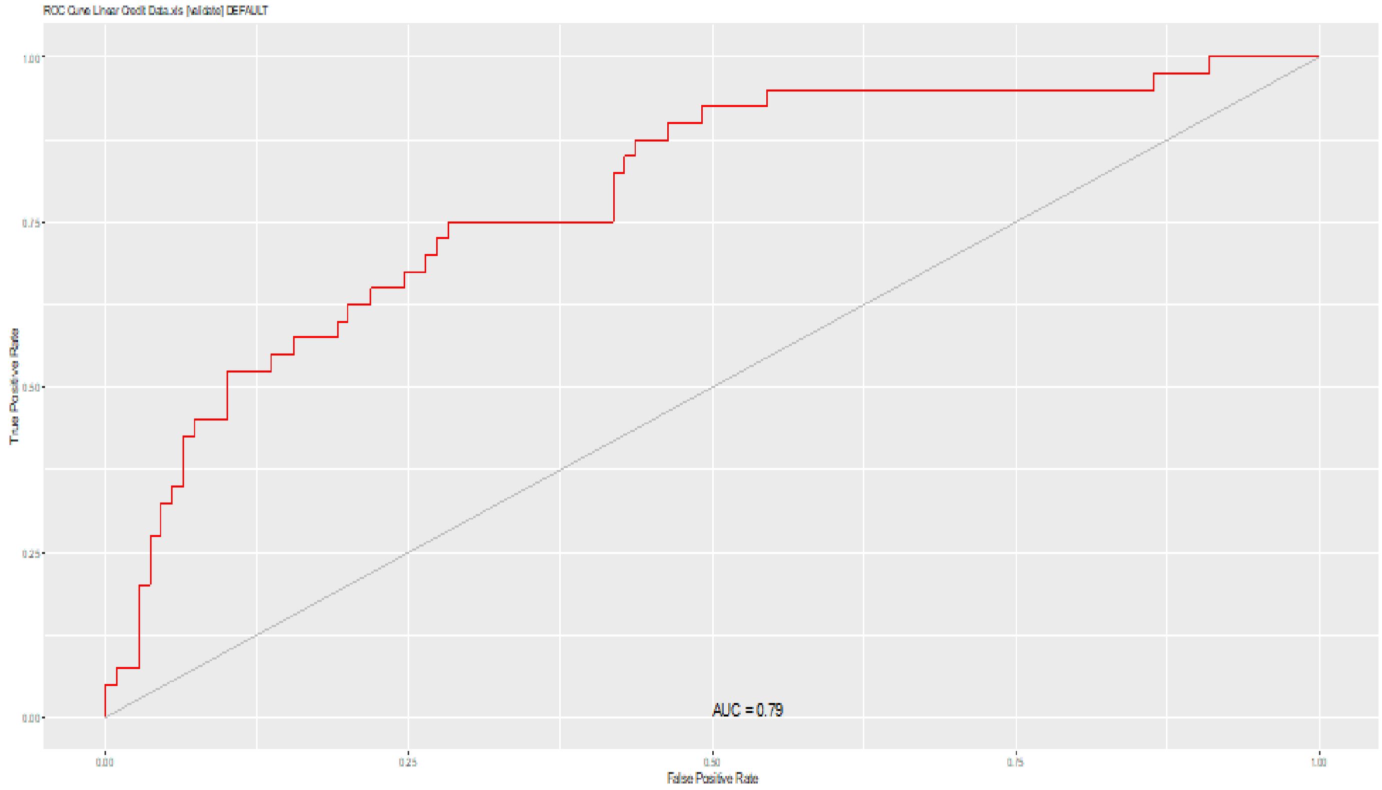
Risk Variable: Report: ○ Class ● Probability Include: ● Identifiers ○ All

```
Error matrix for the Linear model on Credit Data.xls [validate] (counts):

      Predicted
Actual  0  1 Error
     0 96 14  12.7
     1 19 21  47.5

Error matrix for the Linear model on Credit Data.xls [validate] (proportions):

      Predicted
Actual    0    1 Error
     0 64.0  9.3  12.7
     1 12.7 14.0  47.5

Overall error: 22%, Averaged class error: 30.1%

Rattle timestamp: 2019-09-14 19:54:23 likane.druide_snhu
======================================================================
```

Logistic Regression Model Results

- ROC

The area under the ROC curve is 0.7920 which is good (good predictive ability)



Model’s goodness of fit

-Pseudo R2 = 0.5369: the model can account for 53.7% data variability.

- Null deviance: 855.21 on 699 degrees of freedom.

-Residual deviance: 636.27 on 669 degrees of freedom.

- Residual deviance < null deviance: the model is a good fit to the data.

Conclusions

- The predictive model can help GE accurately identity customers likely to default on loan payments in 78% of the time.

- The implementation of the model will help GE stay competitive while limiting its risks.

- The cost of the project is relatively low compared to the benefits.

-Predictive Analytics can be applied to other departments of the company for better efficiency, cost saving, improved customer knowledge and better risk assessment.

- Predictive analytics can benefit several other industries that collect lots of data to improve business processes and make better decisions : oil and gas, retail

Recommendations

-Model can make use of additional customers’ data such as credit scores.

-Model maintenance should be performed.

Personal and Professional Reflection

In today’s world, data is collected at an unprecedented rate. Organizations can leverage data analytics to improve their businesses at different levels (process/operations, cost, HR, supply chain, customer service etc.). However, ethical and security considerations should always be in the back of every data analyst’s mind when dealing with data especially individuals’ data and financial data. For this project, the dataset contained loan applicants’ data, so privacy and security were my concerns. The dataset was anonymized so that’s was a good thing. However; the data sources were not clearly identified but the assumption was that they were documented somewhere else in the department.

The capstone project was a real-life example of how data analytics can be used to solve business problems. The capstone highlighted the different steps that must be completed in order to implement a data analytics initiative within an organization. The milestones were designed to make us think about the different aspects of data handling and modeling, even the ones that may not come up naturally to us such as data sources, quality, ethical considerations, reproducibility, project management. I realized that data preparation was a very important step in any data analytic project, and it was one of the stages I had to work on in my revised analytic plan. One class I took that will certainly be

helpful in my future career is Project management. The program helped me strengthen my writing and communications skills.

During this program I realized that implementing a data analytics plan within an enterprise may not be easy for different reasons, so a good start is to get every stakeholder involved and communication is the key. Also, I think that presenting the actual benefits of a data analytic initiative for each department and the whole company can help bring everybody on board.

I learned a lot during this program, and I found predictive analytics really interesting because of the various applications in so many different industries. The programming courses (especially Java) were the most challenging classes. Now that I have the foundation, I think it will be beneficial if I can strengthen my skills in programming (R, Python) and in SQL also. I also believe in the saying: “Practice makes perfect” and so I should be able to master these skills with more exposure.