Loan Loss Analysis

# Introduction

We attained a dataset from Kaggle that is intended to be used for predicting whether a loan will default or not. The dataset has no context or definitions. All but 1 of the variables are masked with no meaningful description (i.e. f1, f2, f3 …. ). The one variable with only inherent context in its name is “loss” which ranges from 0 to 100.

Given the objective of this exercise and the limitations of the dataset we chose to answer the following questions:

**Do the f2 and f5 variables together affect the value of loss when a loss has occurred?**

**Does f2 affect the value of loss when a loss has occurred?**

**Does f5 affect the value of loss when a loss has occurred?**

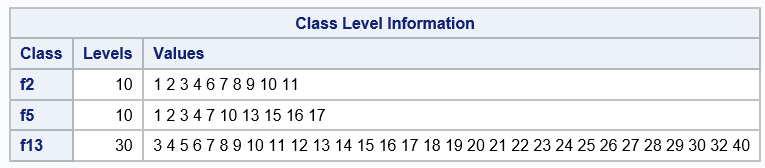
# Exploratory Analysis

The data does not to have any temporal variables therefore we are restricted to a non-time series model selection (ANOVA) to answer the questions of interest.

Because the questions of interest are related only to the value of loss when a loss has occurred we removed all observations from the original data set (of 105200 observations) that had no loss.

For ANOVA we are restricted to the f2 f5 f13 variables which are apparently categorical:

\*f13 may looks continuous however the 30 levels are from 105200 observations, so we treat it as categorical for this study.



# The is clearly interaction as the different levels of f2 and f5; the variances are a concern.

# 

# Model Selection

|  |  |
| --- | --- |
| Forward Selection | Backward Selection |
|  |  |
| Stepwise Selection | LASSO |
|  |  |

Backward Selection has identified that f2 and f5 are the ideal categorical variables for this exercise.

*Not sure what I am doing wrong with LASSO:*

**proc** **glmselect** data = TRAIN plots=all;

class f2 f5 f13;

model loss=f2 f5 f13 / selection = lasso CVDETAILS;

**run**;

**Selected Model:** *loss = B0+B(f2)+B(f5)+B(f2\*f5)*

**proc** **glm** data=train PLOTS=(DIAGNOSTICS RESIDUALS)PLOTS(MAXPOINTS=**10000**);

class f2 f5;

model loss = f2 f5 f2\*f5;

lsmeans f2 f5 f2\*f5 / pdiff tdiff adjust=bon;

**run**;

|  |  |
| --- | --- |
| There is extreme curvature in the residuals | Extreme values causing leverage and “” |
|  |  |
|  |  |

Take log of the loss variable to try to correct residuals

**data** trainlogloss;

set train;

logloss=log(loss);

**run**;

|  |  |
| --- | --- |
| The curvature improved | Improvement but still leverage and “” |
|  |  |
|  |  |

# Final Conclusions

## 

## Means

*Ho: uf2(1) = Uf2(2) = Uf2(….)*

*Ha: uf2(1) <> uf2(2) <> uf2(….)*

There is a clear difference in the means at different levels of f2 (reject Ho) **p value <.0001.**

*Ho: uf5(1) = Uf5(2) = Uf5(….)*

*Ha: uf5(1) <> uf5(2) <> uf5(….)*

There is also evidence to suggest that there is a difference in the means at different levels of f5 (reject Ho) **p value <.048.**

**The f2 and f5 variables do have an effect on the median of loss (p value .01) however it only accounts for .0205 of the variance of median loss. Though statistically significant it is impractical to use f2 and f5 variables to determine or predict the loss.**

**The main effect f2 variable does have an effect by itself while accounting for f5 (p value .0396) however it also only accounts for a small portion of the median loss making it practically insignificant.**

**The main effect f5 variable also does have an effect by itself while accounting for f2 (p value .0136) however it also only accounts for a small portion of the median loss making it practically insignificant.**

**In this observational study we have no information about how the data was collected. We also have no context as to what the f2 and f5 variables actually are therefore no inference can really be drawn at all.**