German Credit Data Analysis

By Krystin Sinclair, Ye-in Jeon and Christian Braz

June 25, 2018

Overview

- Background of Dataset
- Data Exploration
- Classification
  - Frequentist and Bayesian Method
- Regression
  - Frequentist and Bayesian Method
- Conclusion

Data Structure

```
'data.frame':   999 obs. of  21 variables:
 $ Status_Checking       : Factor w/ 4 levels "A11","A12","A13",..: 2 4 1 1 4 4 2 4 2 2 ...
 $ Duration              : int  48 12 42 24 36 24 36 12 30 12 ...
 $ Credit_history        : Factor w/ 5 levels "A30","A31","A32",..: 3 5 3 4 3 3 3 3 5 3 ...
 $ Purpose               : Factor w/ 10 levels "A40","A41","A410",..: 5 8 4 1 8 4 2 5 1 1 ...
 $ Credit_amount         : int  5951 2096 7882 4870 9055 2835 6948 3059 5234 1295 ...
 $ Saving                : Factor w/ 5 levels "A61","A62","A63",..: 1 1 1 1 5 3 1 4 1 1 ...
 $ Empolyment_duration   : Factor w/ 5 levels "A71","A72","A73",..: 3 4 4 3 3 5 3 4 1 2 ...
 $ Installment_rate      : int  2 2 2 3 2 3 2 2 4 3 ...
 $ Personal_status       : Factor w/ 4 levels "A91","A92","A93",..: 2 3 3 3 3 3 3 1 4 2 ...
 $ Otherdebtors          : Factor w/ 3 levels "A101","A102",..: 1 1 3 1 1 1 1 1 1 1 ...
 $ Residence_Year        : int  2 3 4 4 4 4 2 4 2 1 ...
 $ Property              : Factor w/ 4 levels "A121","A122",..: 1 1 2 4 4 2 3 1 3 3 ...
 $ Age                   : int  22 49 45 53 35 53 35 61 28 25 ...
 $ Other_installment_plan: Factor w/ 3 levels "A141","A142",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Housing               : Factor w/ 3 levels "A151","A152",..: 2 2 3 3 3 2 1 2 2 1 ...
 $ Exisiting_credits     : int  1 1 1 2 1 1 1 1 2 1 ...
 $ Job                   : Factor w/ 4 levels "A171","A172",..: 3 2 3 3 2 3 4 2 4 3 ...
 $ Liable_People         : int  1 2 2 2 2 1 1 1 1 1 ...
 $ Telephone             : Factor w/ 2 levels "A191","A192": 1 1 1 1 2 1 2 1 1 1 ...
 $ Foreign_worker        : Factor w/ 2 levels "A201","A202": 1 1 1 1 1 1 1 1 1 1 ...
 $ Credit_default        : Factor w/ 2 levels "1","2": 2 1 1 2 1 1 1 1 2 2 ...
```

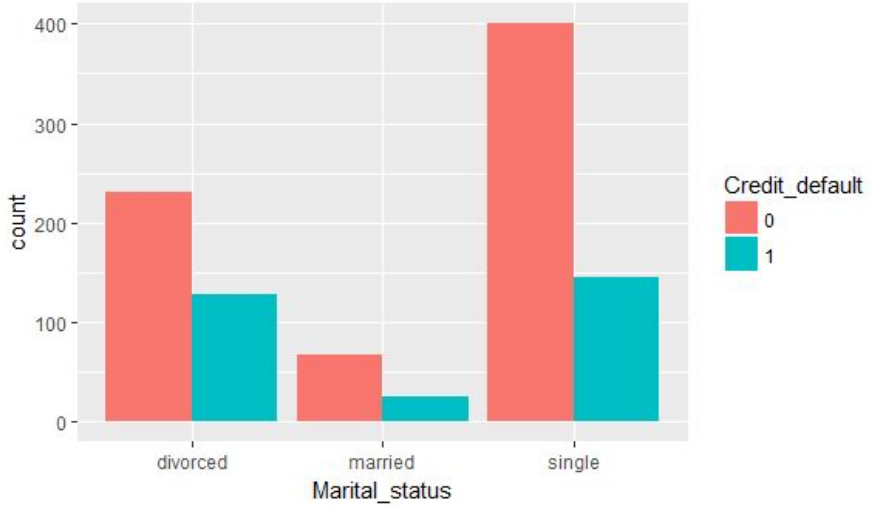
Summary Statistics

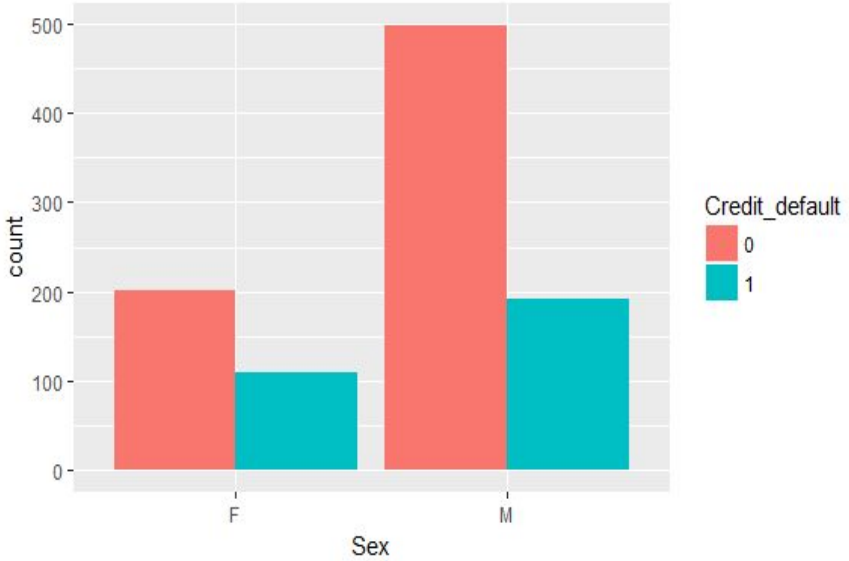
```
Status_Checking    Duration     Credit_history    Purpose    Credit_amount   Saving
A11:273         Min.   : 4.00   A30: 40        A43    :279   Min.   :  250   A61:603
A12:269         1st Qu.:12.00   A31: 49        A40    :234   1st Qu.: 1368   A62:103
A13: 63         Median :18.00   A32:530        A42    :181   Median : 2320   A63: 63
A14:394         Mean   :20.92   A33: 88        A41    :103   Mean   : 3273   A64: 48
                3rd Qu.:24.00   A34:292        A49    : 97   3rd Qu.: 3972   A65:182
                Max.   :72.00                  A46    : 50   Max.   :18424
                                               (Other): 55
Empolyment_duration Installment_rate Personal_status Otherdebtors Residence_Year  Property
A71: 62             Min.   :1.000    A91: 50         A101:906     Min.   :1.000   A121:281
A72:172             1st Qu.:2.000    A92:310         A102: 41     1st Qu.:2.000   A122:232
A73:339             Median :3.000    A93:547         A103: 52     Median :3.000   A123:332
A74:174             Mean   :2.972    A94: 92                      Mean   :2.844   A124:154
A75:252             3rd Qu.:4.000                                 3rd Qu.:4.000
                    Max.   :4.000                                 Max.   :4.000

     Age        Other_installment_plan Housing    Exisiting_credits   Job      Liable_People
Min.   :19.00   A141:139               A151:179   Min.   :1.000     A171: 22   Min.   :1.000
1st Qu.:27.00   A142: 47               A152:712   1st Qu.:1.000     A172:200   1st Qu.:1.000
Median :33.00   A143:813               A153:108   Median :1.000     A173:629   Median :1.000
Mean   :35.51                                     Mean   :1.406     A174:148   Mean   :1.155
3rd Qu.:42.00                                     3rd Qu.:2.000                3rd Qu.:1.000
Max.   :75.00                                     Max.   :4.000                Max.   :2.000

Telephone  Foreign_worker Credit_default
A191:596   A201:962       1:699
A192:403   A202: 37       2:300
```

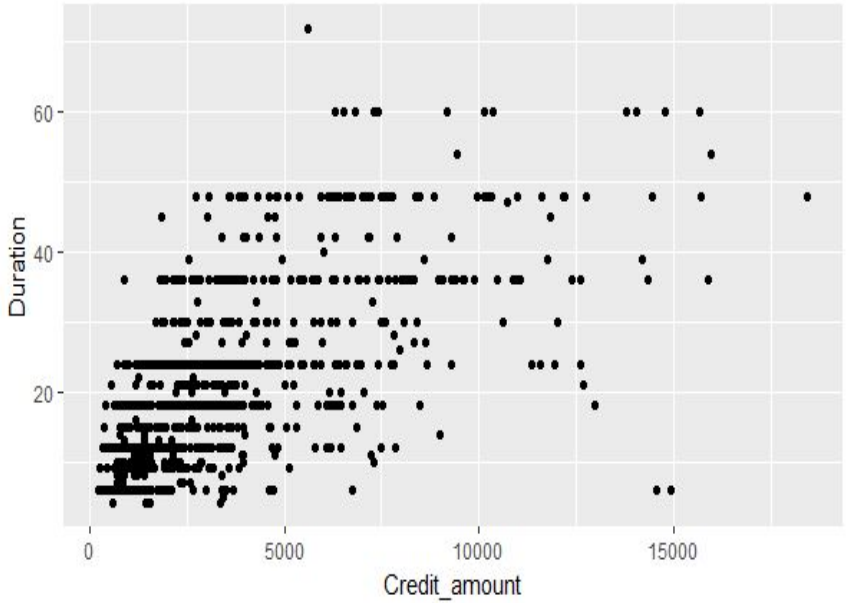
Exploratory Data Analysis-Bar plot

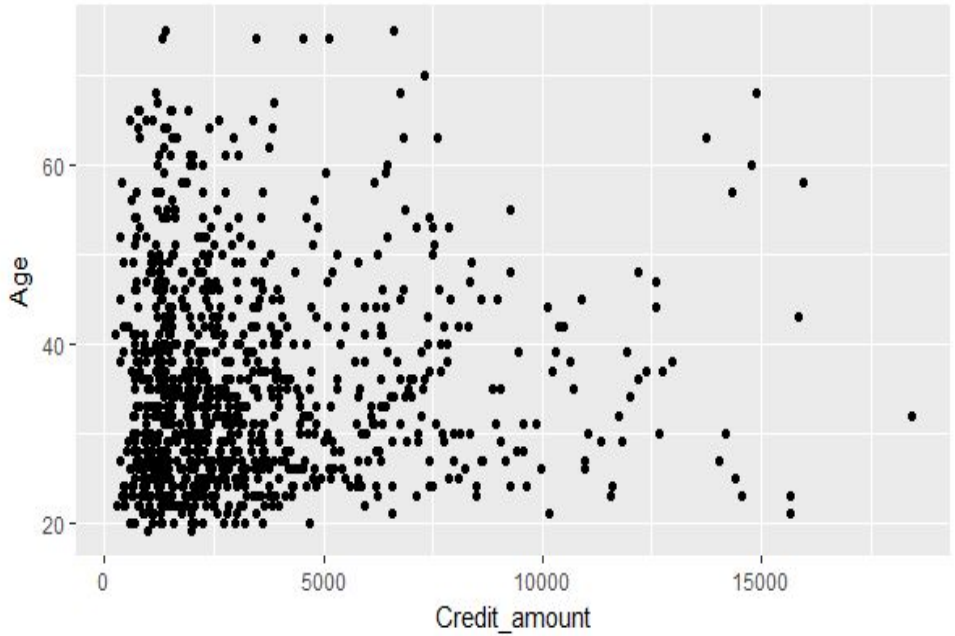




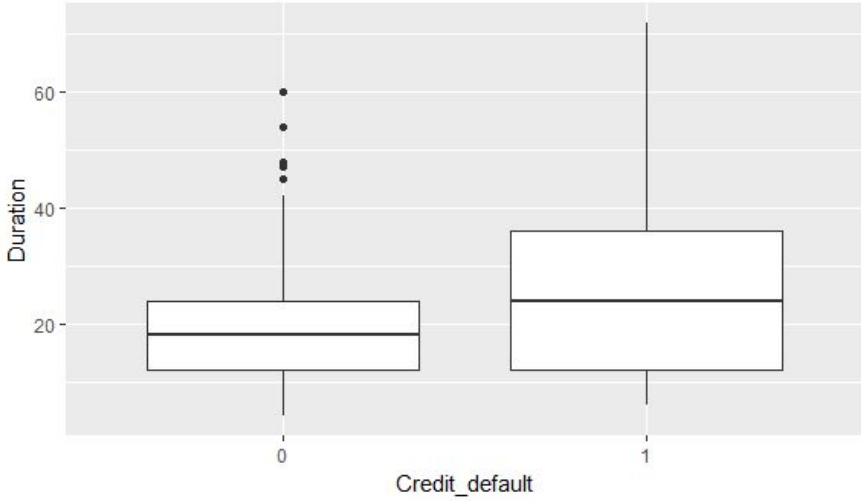
- Marital Status
  - Biggest difference between good and bad credit is in Single applicants
- Gender
  - Majority and men and more than double of the man with bad credit and man with good credit

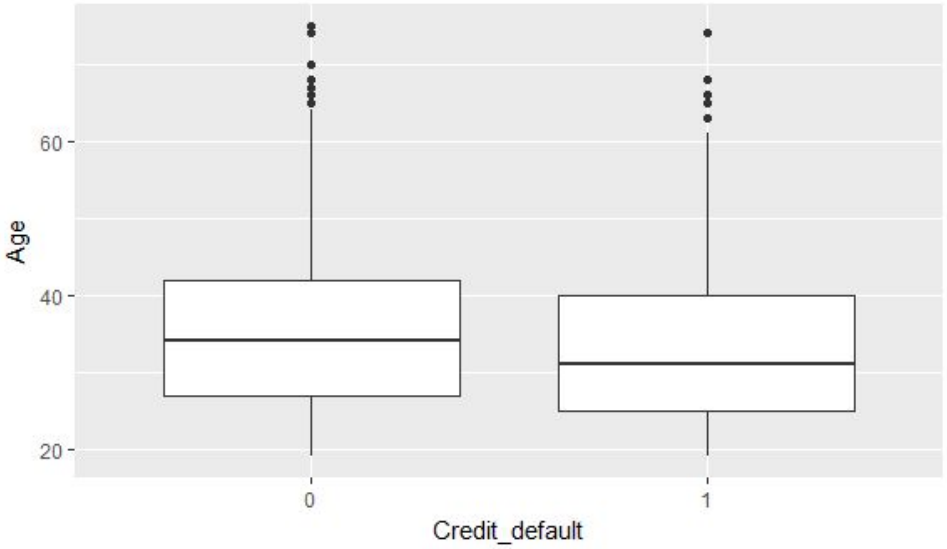
Exploratory Data Analysis-Scatter plot





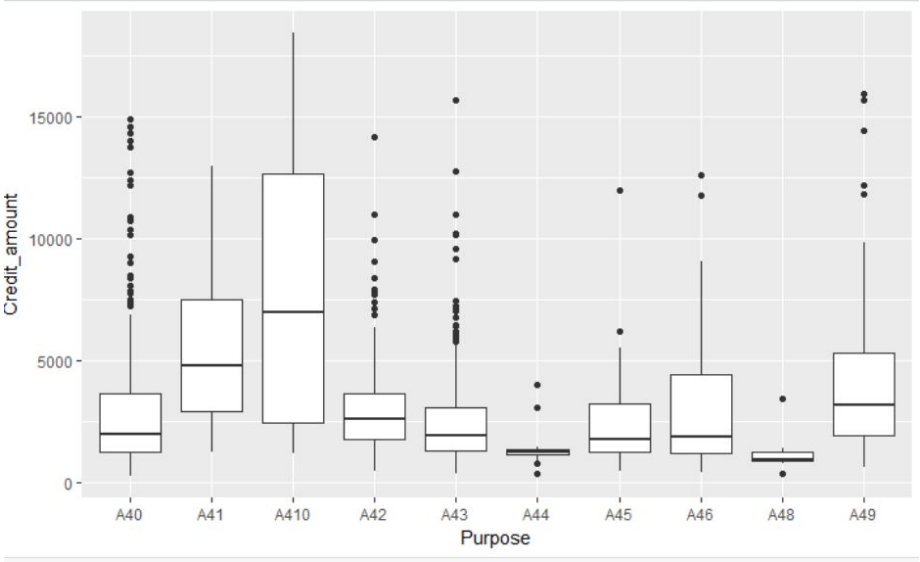
Exploratory Data Analysis-Boxplot

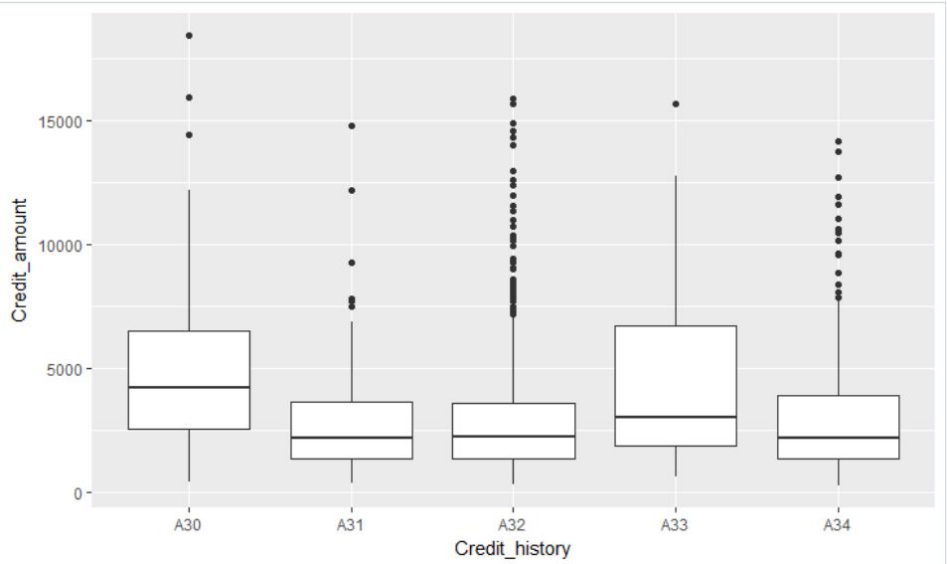




- Duration of the loans in months
  - Boxplot shows Defaulters may have loans of longer time frames
- Age of the loan applicant
  - Boxplot shows no significant difference between Credit Default by Age

Exploratory Data Analysis-Boxplot



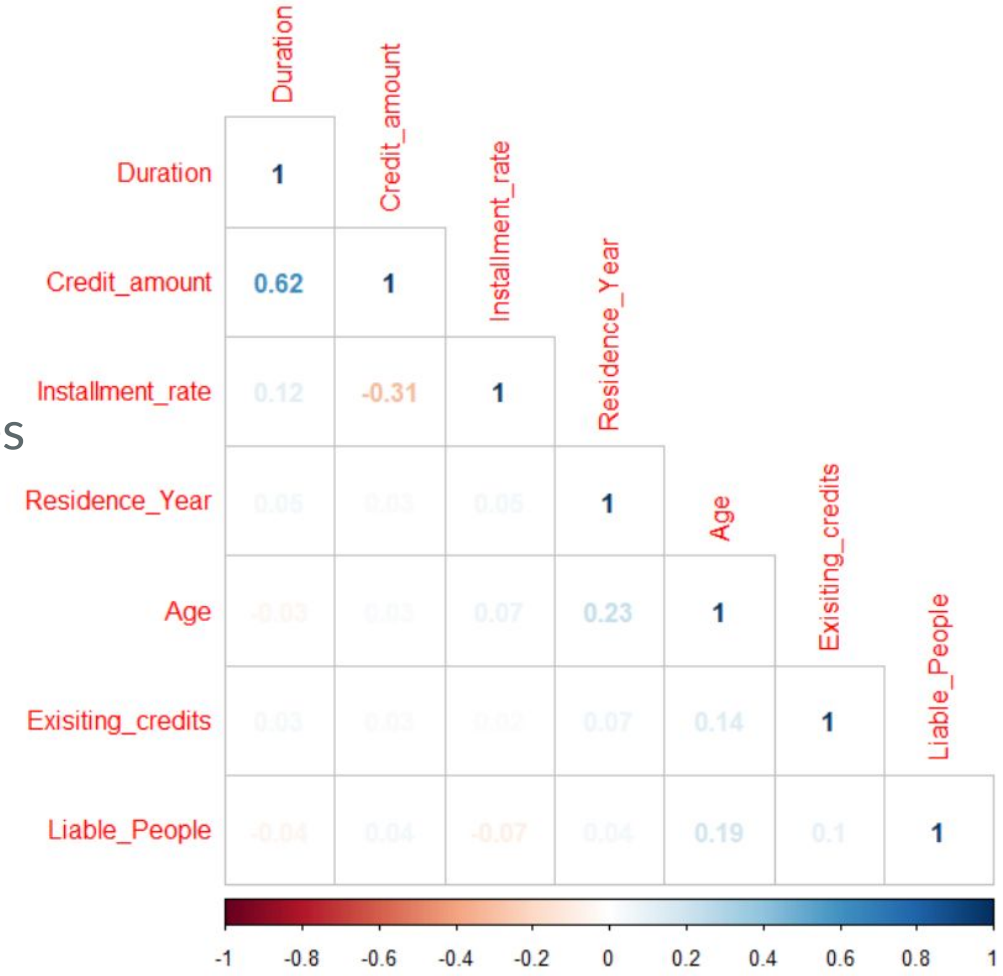


A40: Car(new)
A41: Car(used)
A42: Furniture/equipment
A43: Radio/television
A44: Domestic Appliances
A45: Repairs
A46: Education
A47: Vacation
A48: Retraining
A49: Business
A410:Other

A30 : no credits taken/ all credits paid back duly
A31 : all credits at this bank paid back duly
A32 : existing credits paid back duly till now
A33 : delay in paying off in the past
A34 : critical account/ other credits existing (not at this bank)

Correlation

- Testing for collinearity
  - Minimal correlation between explanatory variables



Regression and logistic regression: feature selection and methodology

- Regression Problem - Understand Credit Amount: How much will a bank give to an applicant?
    - Generalized Linear model - Frequentist and Bayesian Method
        - Use ANOVA to subset data to important attributes
        - Multi-Regression

- Classification Problem - Predict how propense a debtor is in default
    - Generalized Linear model - Frequentist and Bayesian Method
        - Logistic model using a balanced train/test sets

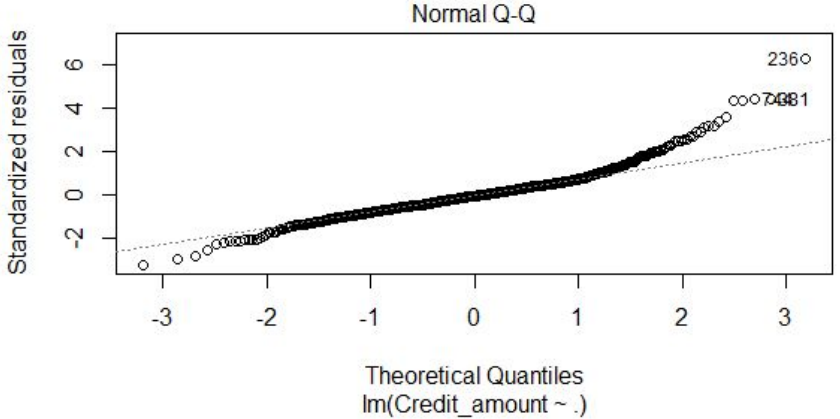
Frequentist Method

Anova to select features

Status Checking
Duration
Purpose
Installment Rate
Property
Job

General Linear Model

Multiple R-squared: 0.6287
Adjusted R-squared: 0.6014
F-statistic: 22.97
p-value: < 2.2e-16



Analysis of Variance Table

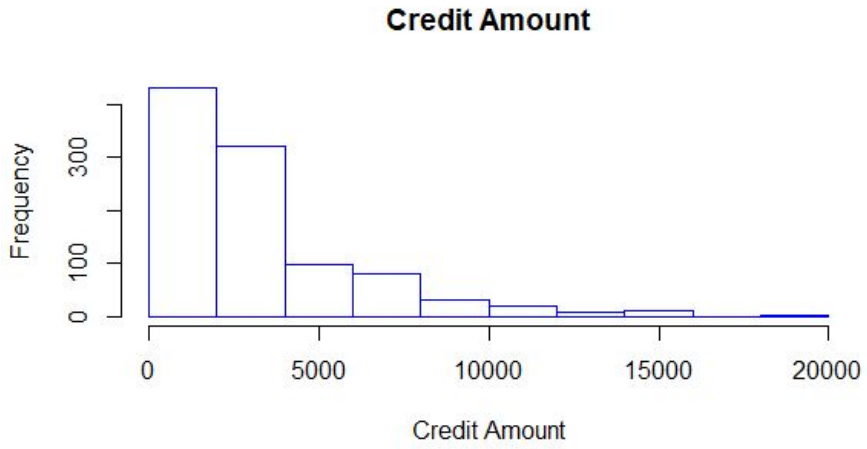
Response: Credit_amount

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Status_Checking	3	102457016	34152339	10.8290	5.976e-07	***
Duration	1	2081080287	2081080287	659.8668	< 2.2e-16	***
Credit_history	4	52351227	13087807	4.1499	0.0025083	**
Purpose	9	420910304	46767812	14.8291	< 2.2e-16	***
Saving	4	49305945	12326486	3.9085	0.0038147	**
Empolyment_duration	4	38017970	9504493	3.0137	0.0176304	*
Installment_rate	1	417534886	417534886	132.3915	< 2.2e-16	***
Otherdebtors	2	22877322	11438661	3.6270	0.0271358	*
Residence_Year	1	3694731	3694731	1.1715	0.2794896	
Property	3	48211936	16070645	5.0957	0.0017154	**
Age	1	454243	454243	0.1440	0.7044299	
Other_installment_plan	2	4586370	2293185	0.7271	0.4836907	
Housing	2	757921	378961	0.1202	0.8867978	
Exisiting_credits	1	398962	398962	0.1265	0.7222016	
Job	3	146943594	48981198	15.5309	9.095e-10	***
Liable_People	1	248922	248922	0.0789	0.7788447	
Telephone	1	45775638	45775638	14.5145	0.0001523	***
Foreign_worker	1	648608	648608	0.2057	0.6503418	
Credit_default	1	6230948	6230948	1.9757	0.1603209	
Sex	1	6024730	6024730	1.9103	0.1674023	
Marital_status	2	28554168	14277084	4.5270	0.0111561	*
Residuals	651	2053116419	3153789			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Frequentist Regression

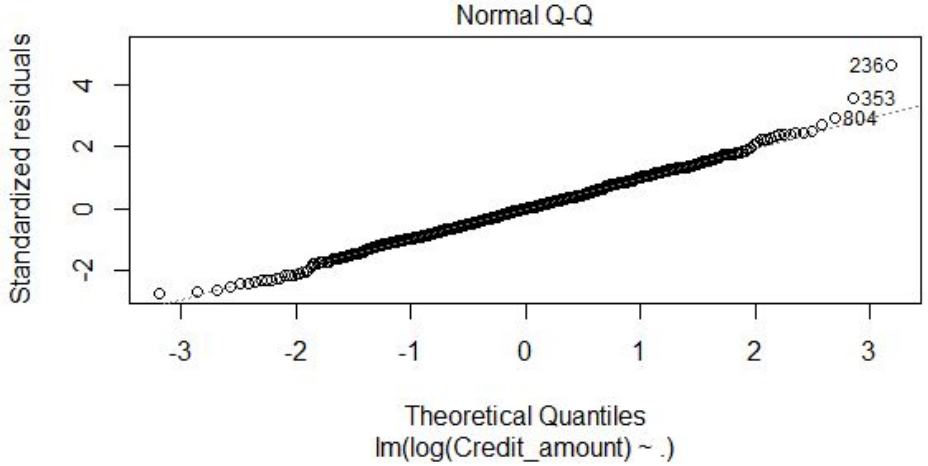
- Credit Amount is skewed
  - use log function to normalize



Regression output

Multiple R-squared: 0.6419,
Adjusted R-squared: 0.6313
F-statistic: 60.86 on 20 and 679 DF,
p-value: < 2.2e-16

Root Mean Squared Error
2080.457



Prediction Example
For each month increase in duration there is 1.04 increase in credit amount.

Bayesian Method

Root Mean Square Error BMA 2026.02574867673

Root Mean Square Error BPM 2019.23719715534

Root Mean Square Error MPM 2042.38372389604

Root Mean Square Error HPM 2035.69450526844

	P(B != 0 | Y)	model 1	model 2	model 3	model 4	model 5
Intercept	1.00000000	1.0000	1.000000	1.000000	1.0000000	1.0000000
Status_CheckingA12	0.05350323	0.0000	0.000000	0.000000	0.0000000	0.0000000
Status_CheckingA13	0.36845303	0.0000	0.000000	1.000000	0.0000000	0.0000000
Status_CheckingA14	0.06662750	0.0000	0.000000	0.000000	0.0000000	0.0000000
Duration	0.99999943	1.0000	1.000000	1.000000	1.0000000	1.0000000
PurposeA41	0.99866581	1.0000	1.000000	1.000000	1.0000000	1.0000000
PurposeA410	0.14543591	0.0000	0.000000	0.000000	0.0000000	0.0000000
PurposeA42	0.95403595	1.0000	1.000000	1.000000	1.0000000	1.0000000
PurposeA43	0.07328262	0.0000	0.000000	0.000000	0.0000000	0.0000000
PurposeA44	0.12699451	0.0000	0.000000	0.000000	0.0000000	0.0000000
PurposeA45	0.04726772	0.0000	0.000000	0.000000	0.0000000	0.0000000
PurposeA46	0.21116943	0.0000	0.000000	0.000000	0.0000000	0.0000000
PurposeA48	0.64107342	1.0000	0.000000	1.000000	1.0000000	1.0000000
PurposeA49	0.29526978	0.0000	0.000000	0.000000	0.0000000	0.0000000
Installment_rate	0.99999619	1.0000	1.000000	1.000000	1.0000000	1.0000000
PropertyA122	0.07447643	0.0000	0.000000	0.000000	0.0000000	0.0000000
PropertyA123	0.20322876	0.0000	0.000000	0.000000	0.0000000	0.0000000
PropertyA124	0.29584713	0.0000	0.000000	0.000000	0.0000000	0.0000000
JobA172	0.25160961	0.0000	0.000000	0.000000	0.0000000	1.0000000
JobA173	0.49705448	1.0000	1.000000	1.000000	0.0000000	0.0000000
JobA174	0.99984035	1.0000	1.000000	1.000000	1.0000000	1.0000000
BF	NA	1.0000	0.740299	0.669354	0.6607175	0.4750494
PostProbs	NA	0.0348	0.025700	0.023300	0.0231000	0.0170000
R2	NA	0.6223	0.618400	0.625400	0.6183000	0.6215000
dim	NA	8.0000	7.000000	9.000000	7.0000000	8.0000000
logmarg	NA	-1794.3533	-1794.653954	-1794.754695	-1794.7676815	-1795.0975891

Marginal Posterior Summaries of Coefficients:

Using BPM

Based on the top 18461 models

	post mean	post SD	post p(B != 0)
Intercept	7.791312	0.017955	1.000000
Status_CheckingA12	0.002068	0.013739	0.053503
Status_CheckingA13	-0.064793	0.096428	0.368453
Status_CheckingA14	0.002944	0.015331	0.066628
Duration	0.038565	0.001628	0.999999
PurposeA41	0.314128	0.068422	0.998666
PurposeA410	0.046561	0.133385	0.145436
PurposeA42	0.170076	0.061251	0.954036
PurposeA43	-0.004304	0.021476	0.073283
PurposeA44	-0.035400	0.111180	0.126995
PurposeA45	0.004047	0.031914	0.047268
PurposeA46	-0.037665	0.084221	0.211169
PurposeA48	-0.602422	0.526819	0.641073
PurposeA49	0.041754	0.073514	0.295270
Installment_rate	-0.249199	0.016552	0.999996
PropertyA122	0.005111	0.026437	0.074476
PropertyA123	0.019372	0.045645	0.203229
PropertyA124	0.037882	0.068383	0.295847
JobA172	-0.026346	0.058276	0.251610
JobA173	0.063038	0.073531	0.497054
JobA174	0.464864	0.084210	0.999840

The Best Model

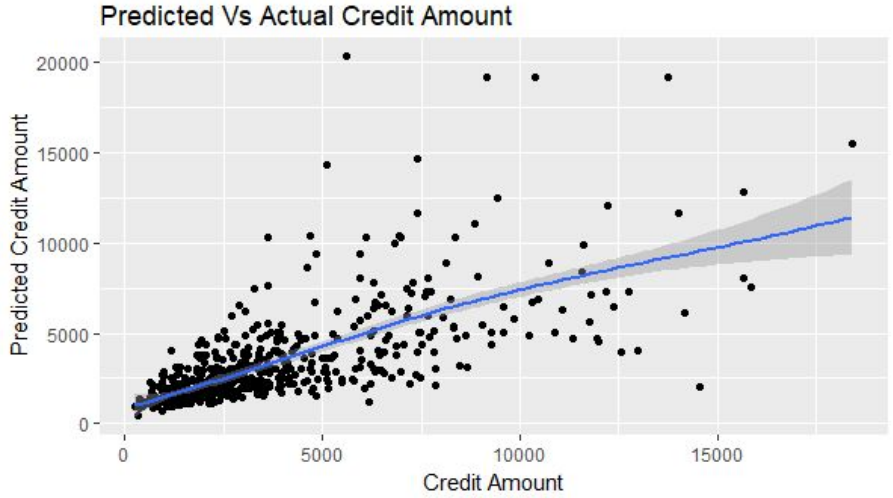
Models

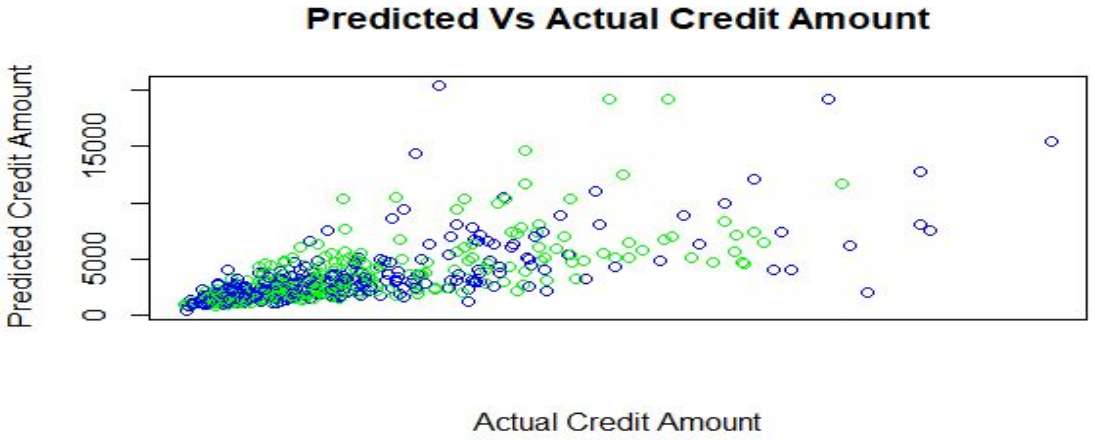
Best Bayesian

RMSE-2019.237

Best Frequentist

RMSE-2080.457





MCMC

Category: Purpose

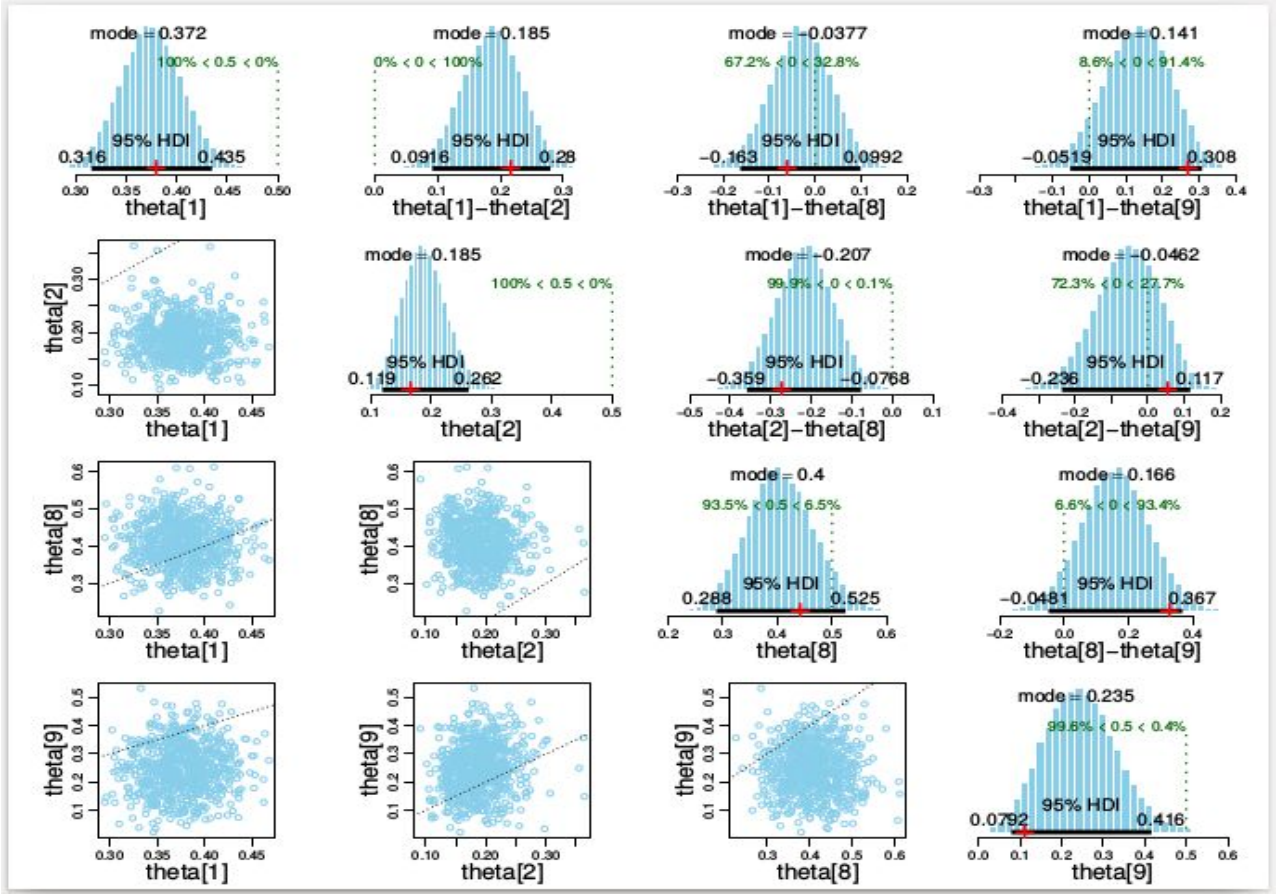
Group Mode: 0.3

1 - Car (new)

2 - Car (used)

8 - Education

9 - Retraining



Logistic Regression Model

- Created a balanced train (70%) and test (30%) sets
- Fit a logistic regression model using the train set
- Predict using the test set
- Metrics
  - Overall Accuracy: 0.735
  - Recall: 0.437
  - Precision: 0.63
  - F-score: 0.52
- For every one month increase in the Duration, the odds of defaulting increase by a factor of 1.02

Bayesian Logistic Regression

- Create a model in JAGS to estimate the coefficients of the linear function of the logistic model
- Use the result of the linear function to make predictions via the sigmoid function (same test set)
- Metrics
  - Overall Accuracy: 0.802
  - Recall: 0.552
  - Precision: 0. 77
  - F-score: 0.64

- Frequentist
  - Overall Accuracy: 0.735
  - Recall: 0.437
  - Precision: 0.63
  - F-score: 0.52

Conclusion

Advice for Bank

- Short term loans are less likely to default
- Older applicants are less likely to default

Advice for applicant

The most important

- Duration
- Purpose
- Property
- Job

Tools

- **R**
  - **Packages**
    - RJAGS
    - GGPLOT2
    - corrplot
    - caret
    - pROC
    - BAS
  - **Charts**
    - Histogram
    - Correlation Matrix
    - Boxplot
    - Scatterplot

References

Data

http://home.cse.ust.hk/~qyang/221/Assignments/German/

Articles

https://loans.usnews.com/beyond-credit-scores-factors-that-affect-a-loan-application

https://studentloanhero.com/featured/personal-loan-purpose-happens-change/

https://www.cbsnews.com/news/5-things-that-can-torpedo-your-mortgage-application/

http://www.ijbf.uum.edu.my/images/pdf/5no1ijbf/6ijbf51.pdf

https://www.sciencedirect.com/science/article/pii/0883902688900183

https://dl.acm.org/citation.cfm?id=131259

https://www.emeraldinsight.com/doi/pdfplus/10.1108/eb013696

http://www.rcmloan.com/credit-building-solana-beach-ca/

Questions?