How property and family affect people's loan application?

Runqi(Ricky) Zhao

12/7/2020

1.Abstract

My project used the dataset from Kaggle to explore how property and working occupations status will affect the evaluation of loan risk, also family status to the loan risk. I picked variables by the descriptions and divided them into 4 parts for the exploratory data analysis and then build multilevel models for family and property sections.

2.Introduction

When people going to apply for loans, we need to submit all kinds of documents to provide that we have the ability to pay the money back later. These documents not only include the properties people already have, but also the information of the income source - working occupation. People with different occupations may have different risk evaluate results about their applications.

During my exploratory data analysis, a variable comes to my notice. It is interesting to know that except all kinds of documents; the loan companies also record people's accompany situation when they submit their applications. So there comes the second question for me: will people's family situation affect their applications?

The dataset comes from Kaggle: https://www.kaggle.com/c/home-credit-default-risk/data. This dataset has more than 300,000 application records, and 122 variables in its train file. I took this file as my raw data.

3.Method

1) Variables selection and Missing Values

At the very beginning, I looked at the variables' descriptions, picked up the factors about:
- **Loans**: Target, Contract Type, Credit Amount, Annuity amount, Price of goods
- **Appliers**: Gender, Education
- **Family**: Accompany, Family Status, Number of children, Number of family members
- **Property and work**:Car, Realty, Total Income, Income Type, Housing Type, Organization Type, Occupation Type

All variables except Occupation Type have small number of missing values (less than 0.5%), so I deleted them. But the occupation Type has a 1/3 missing values. I still want to keep this variable, to make full use of the information from the data, I decided to delete the rows with missing value under four subsets instead of the whole dataset.

2) Exploratory data analysis

First look at the application numbers of and the average risk grades of different occupation types(Figure 1). I noticed that there is difference between different working tpyes, and for the ***Low Skill Labors*** has lower application numbers but has the highest average grade.

Then comes to take a look at the risk grades of different Family Status and Accompany Situations. (Figure 2)

3) Fitting models

In my models, ***Target*** a binomial response variable, 1 indicates that the loan company treat this application as a risk one, while 0 is not risk.

With a binomial response variable, and with category predictors, I took logistic mixed-effect model for this data.

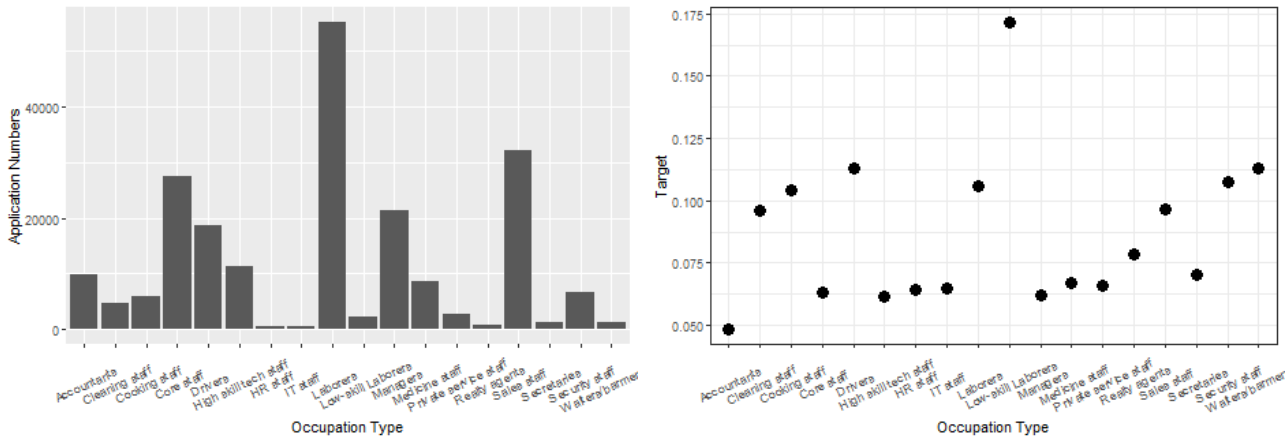

Figure 1: Application Numbers and Target under different Occupation Type

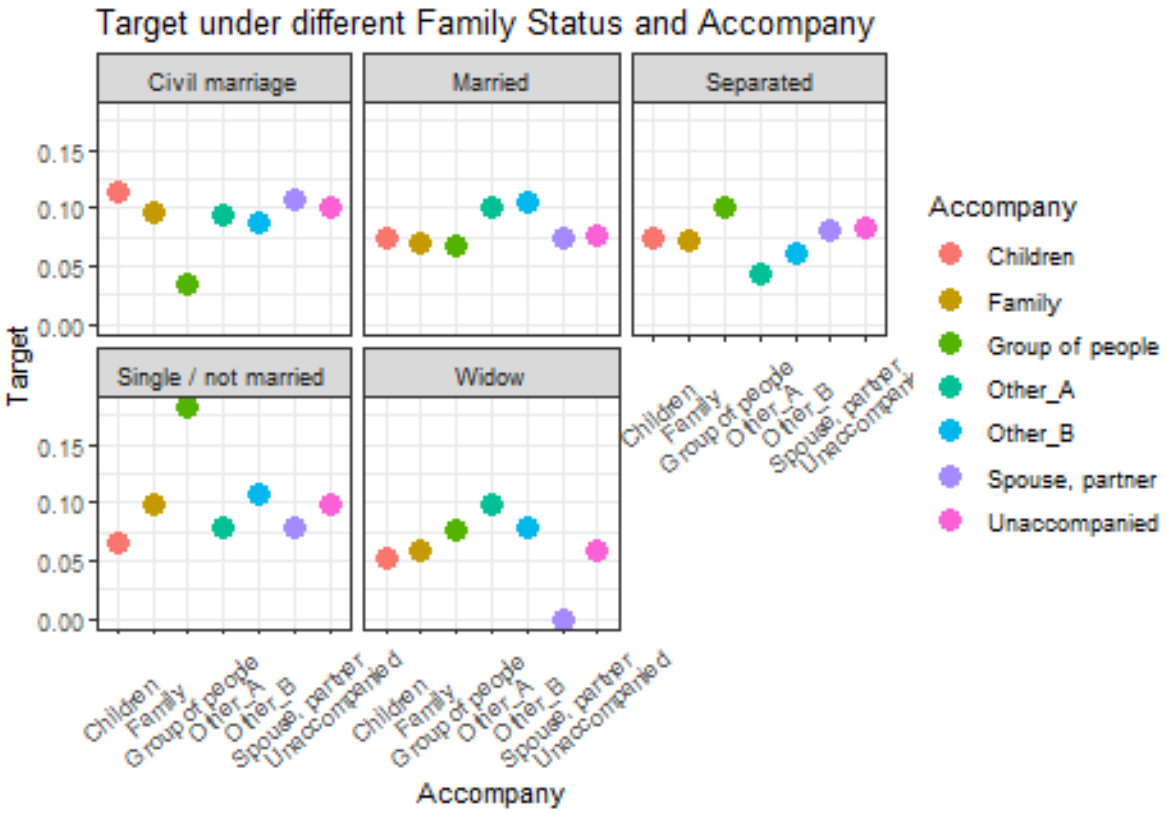

Figure 2: Target of different Family Status and Accompany

For property part, ***Car***, ***Realty***, ***Housing_Type*** are the fixed effect factors indicate the property situation of the applier, and ***Occupation_Type*** is the random effect factor to show appliers' occupations.

The model can be write as:

$Pr(Target_i = 1) = logit^{-1}(\alpha_{j[i]} + \beta^{Car} * Car + \beta^{Realty} * Realty + \beta^{HousingType} * HousingType)$**, for** $i = 1, ..., n$

$\alpha_j \sim$ N$(\mu_\alpha, \sigma^2_{OccupationType})$**, for** $j = 1, ..., 18$

For family part, ***Family_Status*** is the fixed effect factors and ***Accompany*** is the random effect factor to show people's accompany situation when they submit their application.

The model can be write as:

$Pr(Target_i = 1) = logit^{-1}(\alpha_{j[i]} + \beta^{FamilyStatus} * FamilyStatus)$**, for** $i = 1, ..., n$

$\alpha_j \sim$ N$(\mu_\alpha, \sigma^2_{Accompany})$**, for** $j = 1, ..., 7$

4) Model Checking

To fit with *stan_glmer*, I took a subset of the whole dataset, sampled for 10,000 recordings randomly. To check the fitted models, plot the residuals for these two models(Figure 3).

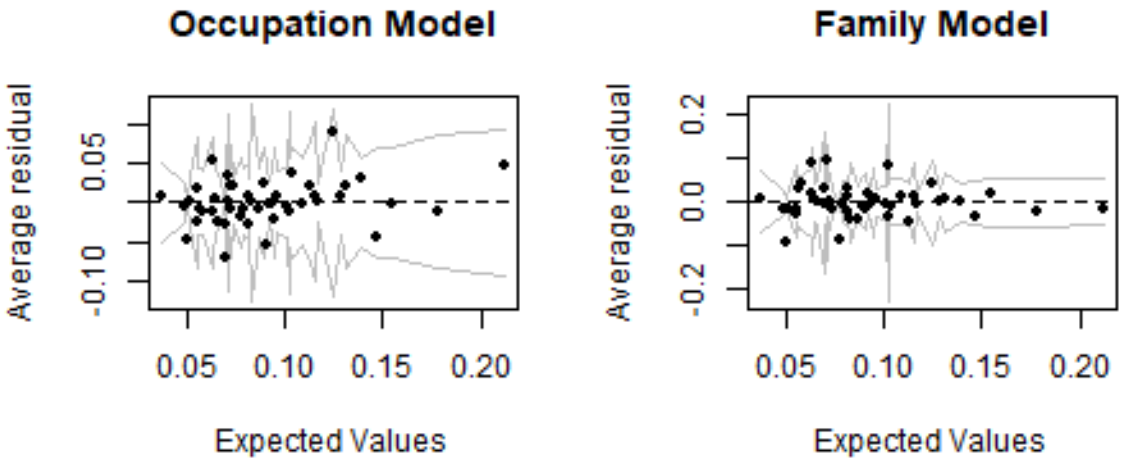

Figure 3: Binned Residual Plots

They look fitting well. Almost all of the points are in the line.

4. Results

First, look at the estimate results of Occupations models(Figure 4).

Car and Realty: The estimated coefficients for people with no realty and no car is 0.3, gives a rough estimate that these guys are 7.5% more likely to be marked as risk applier than people with both of realty and car, for the people live in the same housing type. And people with no car but own realty have a 5% more risk. It is interesting to find that people with car but without realty has 2.5% less likely to be evaluated as risk!

Housing Types: All the estimates are compared with Co-op apartment, this tells that people live in all other housing types are more likely to be treated as risk appliers, under the same car and realty condition. Rented apartment is the most risky type, has 52.5% higher possible, followed by people live with parents, 47.5%, and house or apartment, 35%, then municipal apartment 32.5%.

After controlling for car, realty and housing type, we can look at the effects of different occupations.
Form the estimate results, we can find that drivers and Low skill laborers are the occupations with highest risk, this

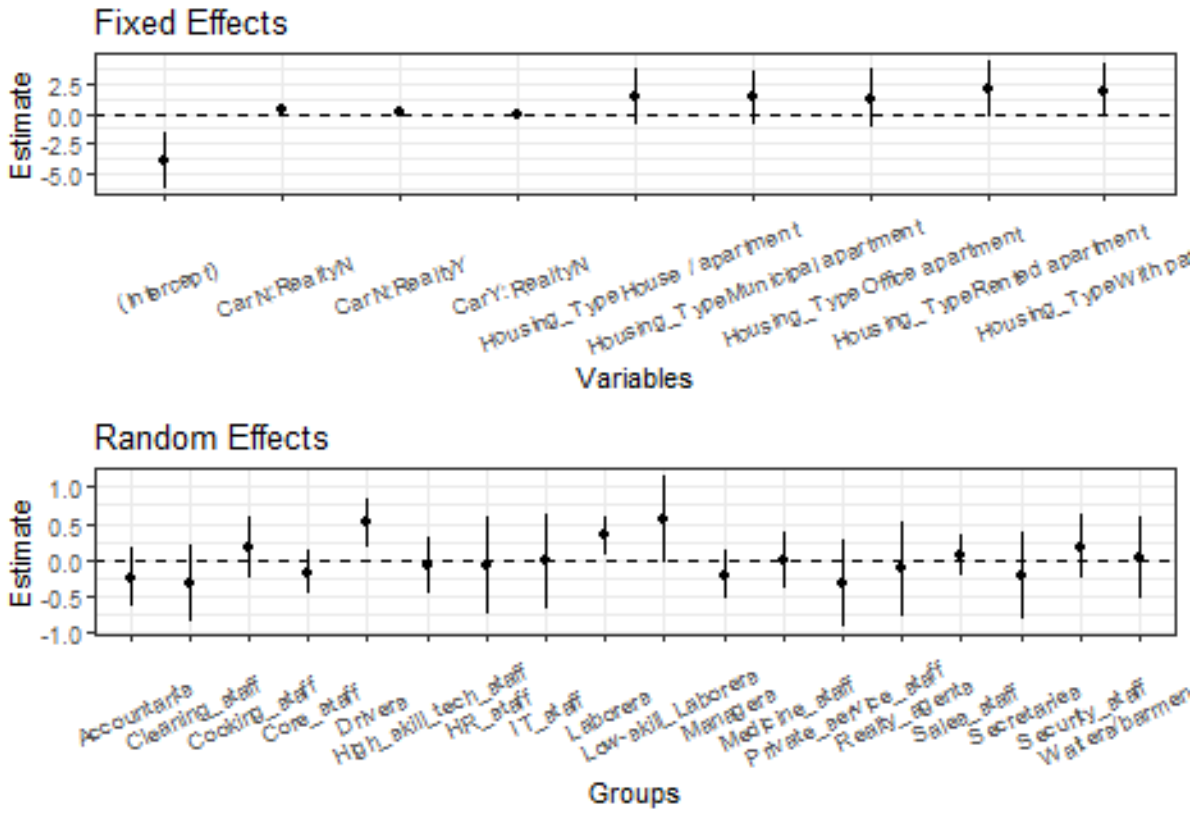

Figure 4: Estimates for Occupations Model

looks consist to the plot of target grade. While Cleaning staff, Private service staff and Secretaries are the lower risk people.

Then look at the estimate results of family models(Figure 5).

Family Status: From the estimate results, we can find that civil marriages have the same risk possible with single or not married people. Compared with them, people who are married get 10% less likely to be treated as risk, separated people have 5% less possible and widow have 27.5%.

For the accompany, we can find that there is no obvious difference between accompanies.

5. Discussion

First, from the models' results, ***the occupations showed different effects on their risk evaluation, while accompany does not.*** Also, we can guess that loan company is ***more welcome Cleaning staff, Private service staff and Secretaries who live in co-op apartment, own their car. Also company may more open to married people and widow.***

With the exploration analysis, we can find the **Low skill laborers** have highest average grade, the drivers have the second. This is consist to our model results. But the lowest average grade, the **Accountants**, does not show a competitive strength in the model. Also, from the plot of target under different Family_Status and Accompany, there shows difference between groups, but the model does not give a variety.

One of the reasons is that the model is fitted by a subset, so there are some difference. To look at this, I may try different sample seed for the subset in the future.
If the models show the similar results, we can dig into some specific occupation types to look into the reasons.

For the family results, one question for me is the difference between civil marriages and married. They have meaningful difference in the results, but to my understanding, they are all married. There may be some difference between these two kinds of marriage, and I can go for more materials for the reason.

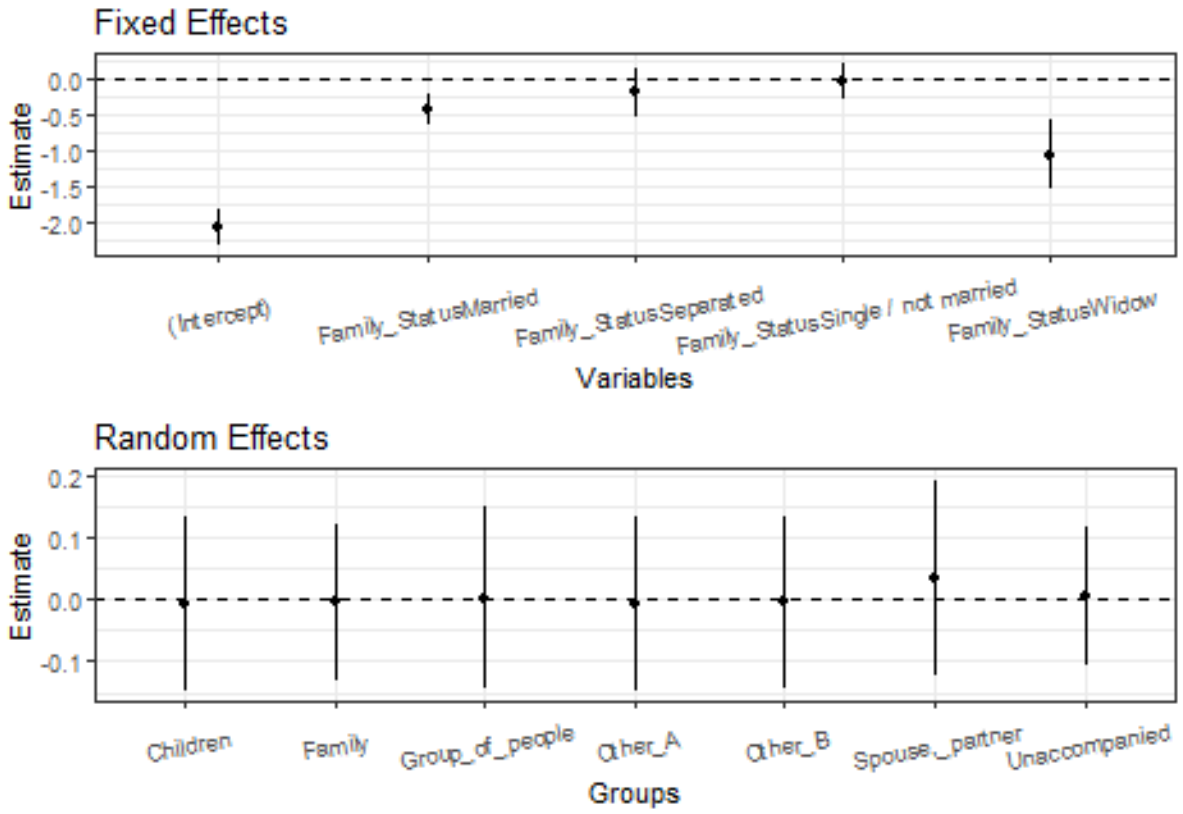

Figure 5: Estimates for Family Model

Another thing to notice is that widow are more likely to be treated as low risk, so there is a rank as $Window < Married < Separated < civilmarriage = Single/NotMarried$. So I may give a guess that this result may be caused by the age and savings. Generally, single people have a smaller average age while they do not have good saving habit, while married people usually have stable income and save more for their family. But all these guesses need to be proved by other analysis.

Appendix

1. EDA Plots not included in the report

1) About Loan

2) About Family

3) About House Type

2. Modeling Plots and results not included in the report

1) Plot to comapre the distribution of data and simulated result(Occupations Model)

2) Plot to comapre the distribution of data and simulated result(Family Model)

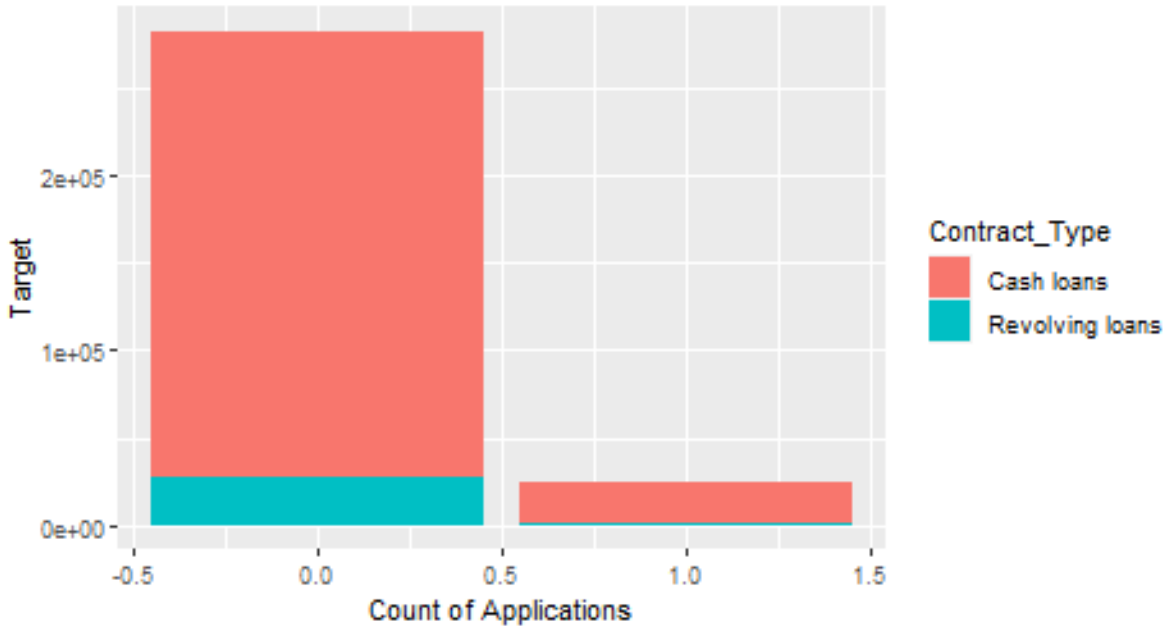

Figure 6: Applications of different Contract Type

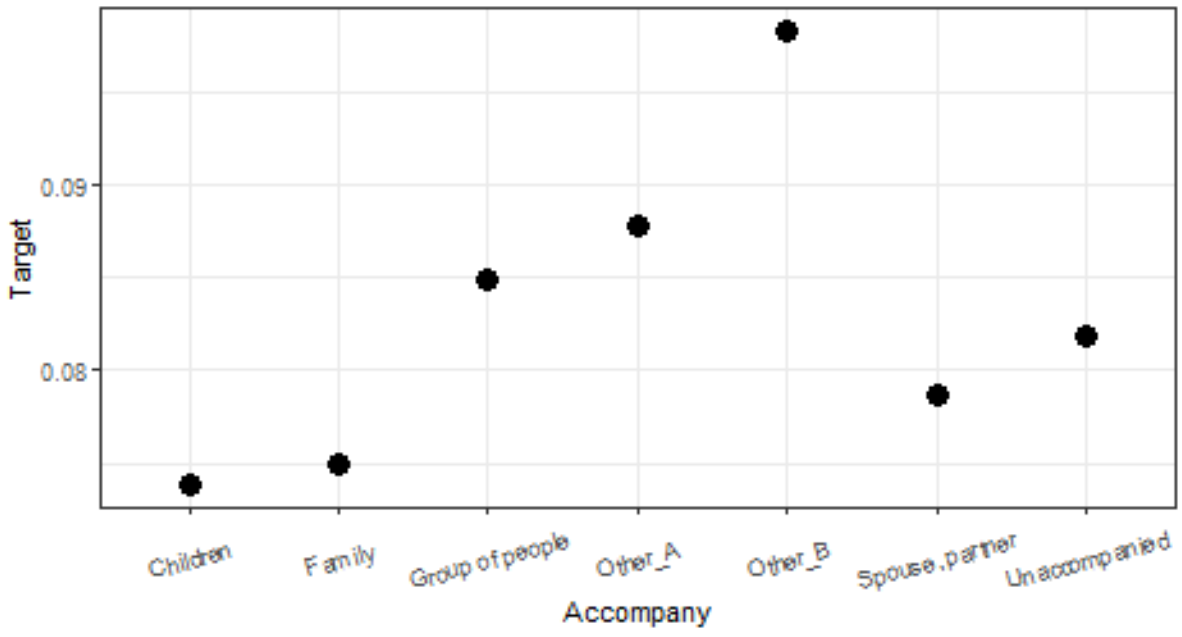

Figure 7: Target of different Accompany

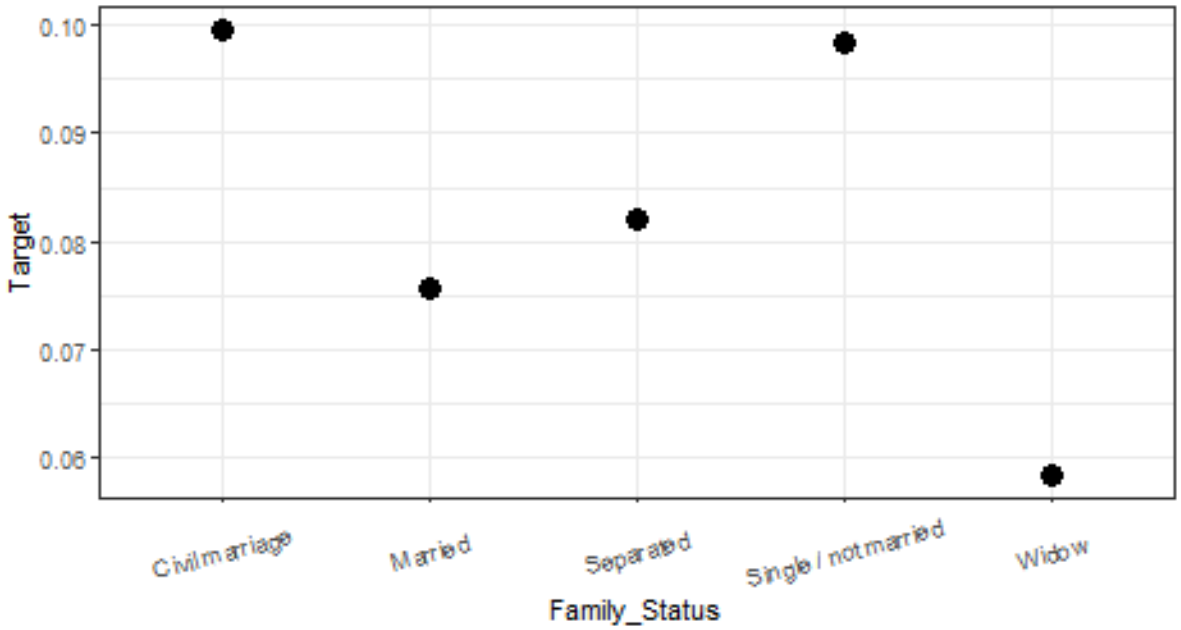

Figure 8: Target of different Family Status

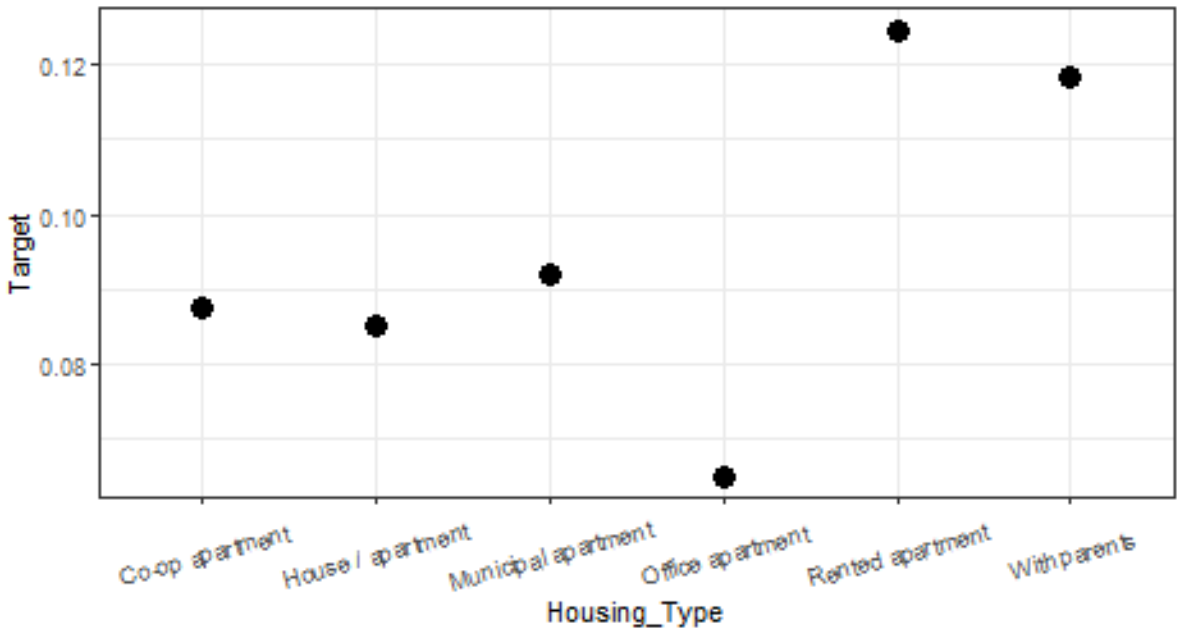

Figure 9: Target of different Housing Type

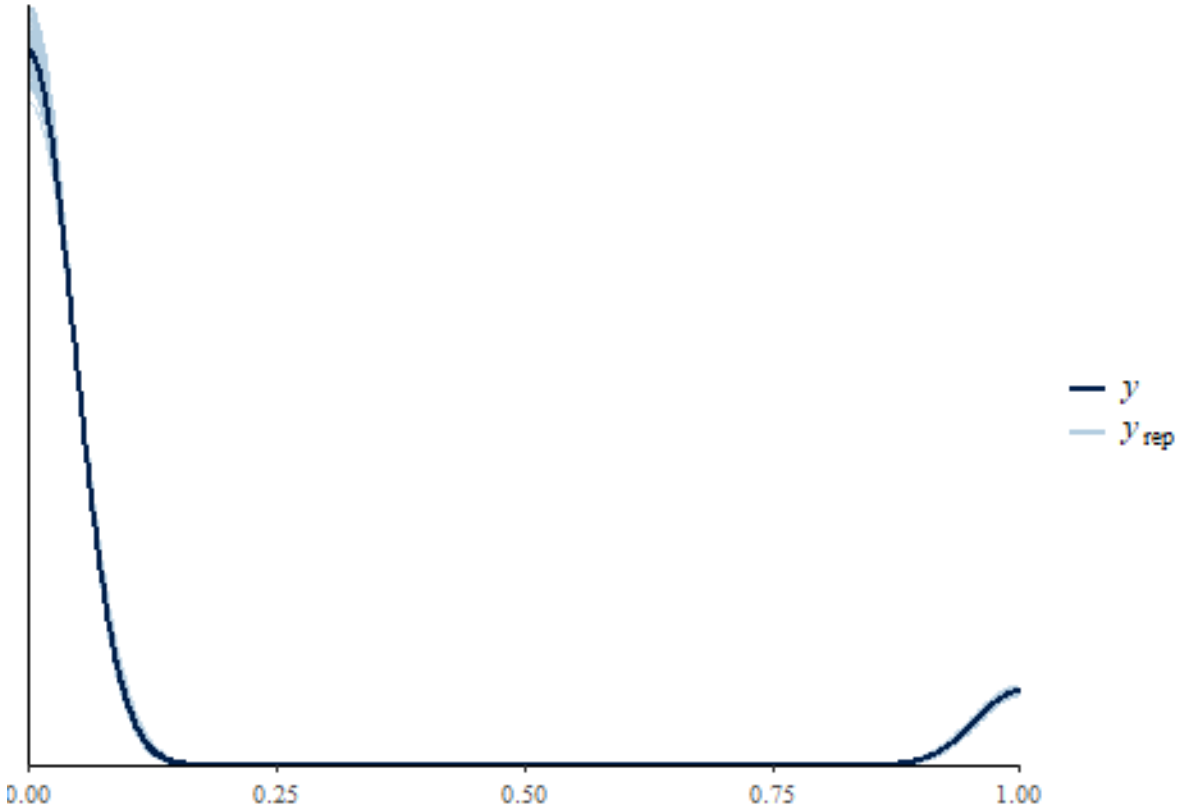

Figure 10: Fitted Polt for Occupations Model

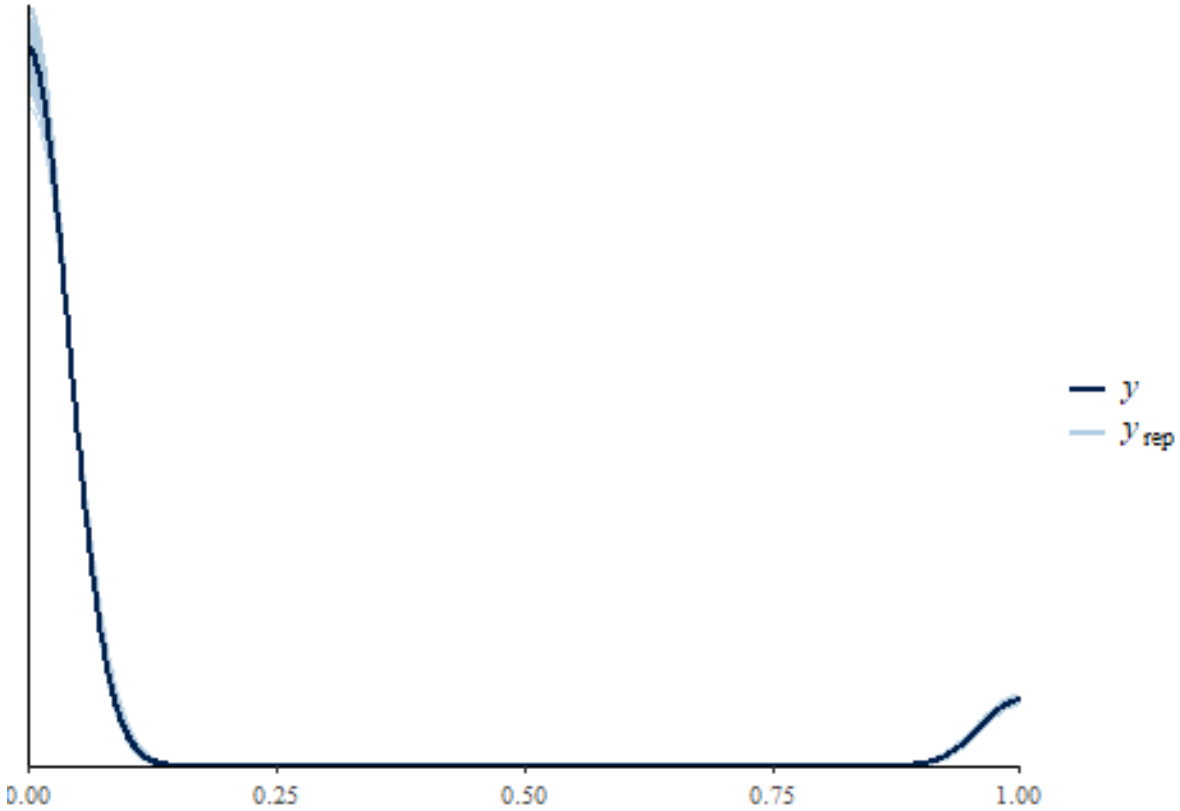

Figure 11: Fitted Polt for Family Model