SF2930 Regression Analysis VT23
Project 2 Report

Xindi Liu(xindi@kth.se)

March 6, 2023

1 Risk arguments

I chose the following risk arguments: `NumberOfFtEmployees`, `CompanyAge`, `TravellingArea`, `NumberOfPersons`, `FinancialRating`, `ActivityCode`. The reason to not chose `DangerousAreas` is because there are too little data in the `Not excluded` category. Other columns are used in the computation of responces therefore they cannot be used as regressor.

2 Grouping

NumberOfPersons: I think the default grouping works well.

ActivityCode: In the default grouping there are too little data in the `Industry` group. I moved some reasonable activities from `Other` to `Industry` or `Service`. I tried to move `Missing` and `X` to an independent group, but this group has too little data. So I merged it with `Industry` because I see from the plot that these two groups have very close values in frequency and severity.

CompanyAge: I first plotted the frequency and severity without grouping, then I see that there are clearly different patterns between the two sides of `CompanyAge=60`. I also found other good edge cuttings at `CompanyAge=10` `CompanyAge=30`.

NumberOfFtEmployees: Through watching the confidence intervals and the group plots I observed 30, 60 and 100 are good edges.

TravellingArea: The difference between home country and nordit is small. Therefore I merged these two groups.

FinancialRating: It can be observed that missing and AAA are similar, and from AA to B are similar. Then I grouped the rest in one group.

3 Model selection

First I tried to add more variables on the default model with the two variables `ActivityCode` and `NumberOfPersons`, and compared their AIC and log-likelihood with the default model. The result is in Table 1.

	Frequency model		Severity model	
	AIC	Log-Likelihood	AIC	Log-Likelihood
Default	170	-77	138	-59
Default + CompanyAge	444	-211	313	-145
Default + FinancialRating	325	-152	237	-108
Default + NumberOfFtEmployees	323	-150	236	-107
Default + TravellingArea	250	-114	178	-79

Table 1: Change of AIC and log-likelihood when adding variables in the default model

The result shows that adding variables does not improve the model.

Then tried other combinations of two variables. The AIC, BIC and Gini scores are presented in Tabel 2, 3, 4.

	Activity	CompanyAge	Financial	NOFtEmployees	NOPersons
CompanyAge	99 100				
Financial	62 68	64 85			
NOFtEmployees	105 91	158 120	63 60		
NOPersons	169 134	171 156	122 123	143 122	
Travelling	53 44	55 55	51 55	53 35	97 87

Table 2: AIC of all models with 2 variables. In each cell, the upper value is for the frequency model, the bottom value is for the severity model.

	Activity	CompanyAge	Financial	NOFtEmployees	NOPersons
CompanyAge	14 31				
Financial	-3 -0.2	-14 9			
NOFtEmployees	19 5	56 58	-11 -4		
NOPersons	33 0.5	8 23	2 24	7 41	
Travelling	-6 -4	-12 3	-4 -2	-10 -1	-14 -7

Table 3: BIC of all models with 2 variables. In each cell, the upper value is for the frequency model, the bottom value is for the severity model.

	Activity	CompanyAge	Financial	NOFtEmployees	NOPersons
CompanyAge	0.226133				
Financial	0.421786	0.356537			
NOFtEmployees	0.191981	0.216777	0.337339		
NOPersons	0.639891	0.634687	0.648897	0.547328	
Travelling	0.308768	0.280691	0.431943	0.187183	0.607617

Table 4: Gini scores of all models with 2 variables.

There are other models that have lower AIC or BIC than the default value. But there Giniscore is much lower than the default value. So I don't like these models.

The combination of `FinancialRating` and `NumberOfPersons` has both higher Giniscore and lower AIC and BIC than the default model, so I think this model is better than the default model.

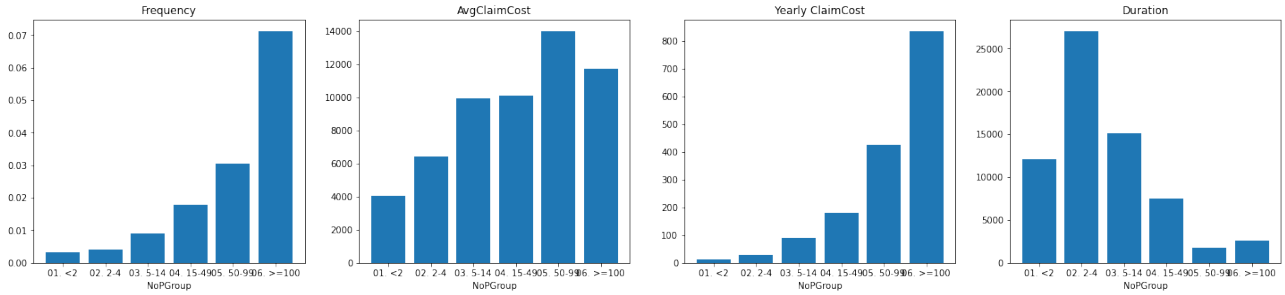
I also tested adding other variables to this model but it does not improve the model.

4 Result

4.1 Grouping

4.1.1 NoPGroup

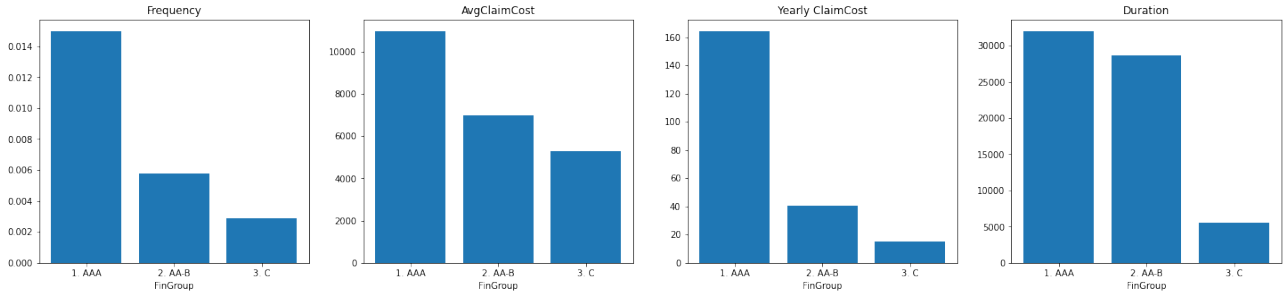
NoPGroup	NumberOfRows	Duration	NumberOfClaims	ClaimCost
01. < 2	25661	12144.287948	40	162369.00
02. 2-4	56200	27041.206917	113	725059.74
03. 5-14	31884	15111.614485	137	1362309.36
04. 15-49	15656	7492.322449	134	1357217.12
05. 50-99	3527	1774.073531	54	755934.00
06. >= 100	3965	2554.897724	182	2131658.61



4.1.2 FinGroup

```
    grouping_map = {
    "AAA" : "1. AAA",
    "AA" : "2. AA-B",
    "A" : "2. AA-B",
    "B" : "2. AA-B",
    "C" : "3. C",
    "AN": "3. C",
    "IR": "3. C",
    "Missing": "1. AAA",
}
```

FinGroup	NumberOfRows	Duration	NumberOfClaims	ClaimCost
1. AAA	65747	31996.419440	479	5254903.17
2. AA-B	59410	28609.525129	165	1155072.16
3. C	11736	5512.458485	16	84572.50



4.2 Confidence intervals

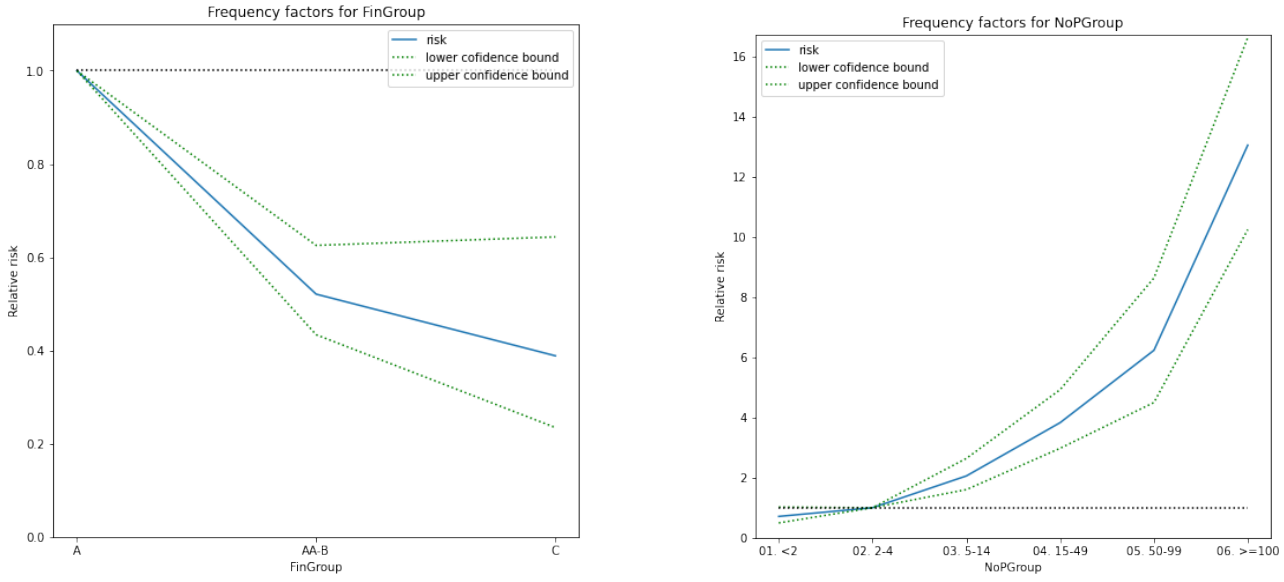


Figure 1: Three simple graphs

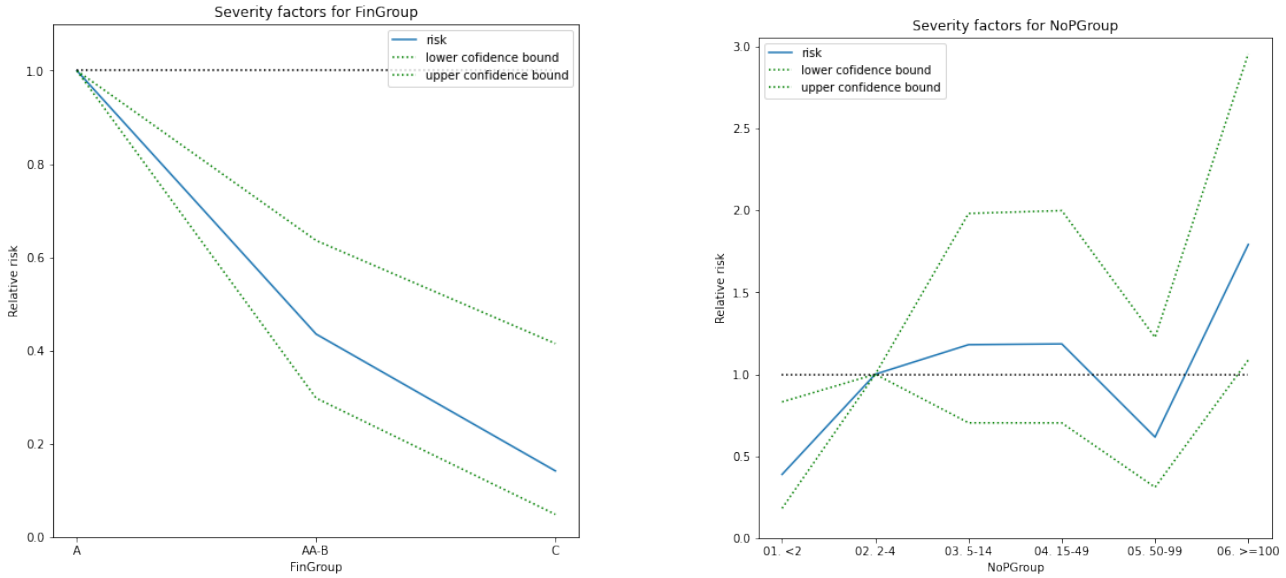
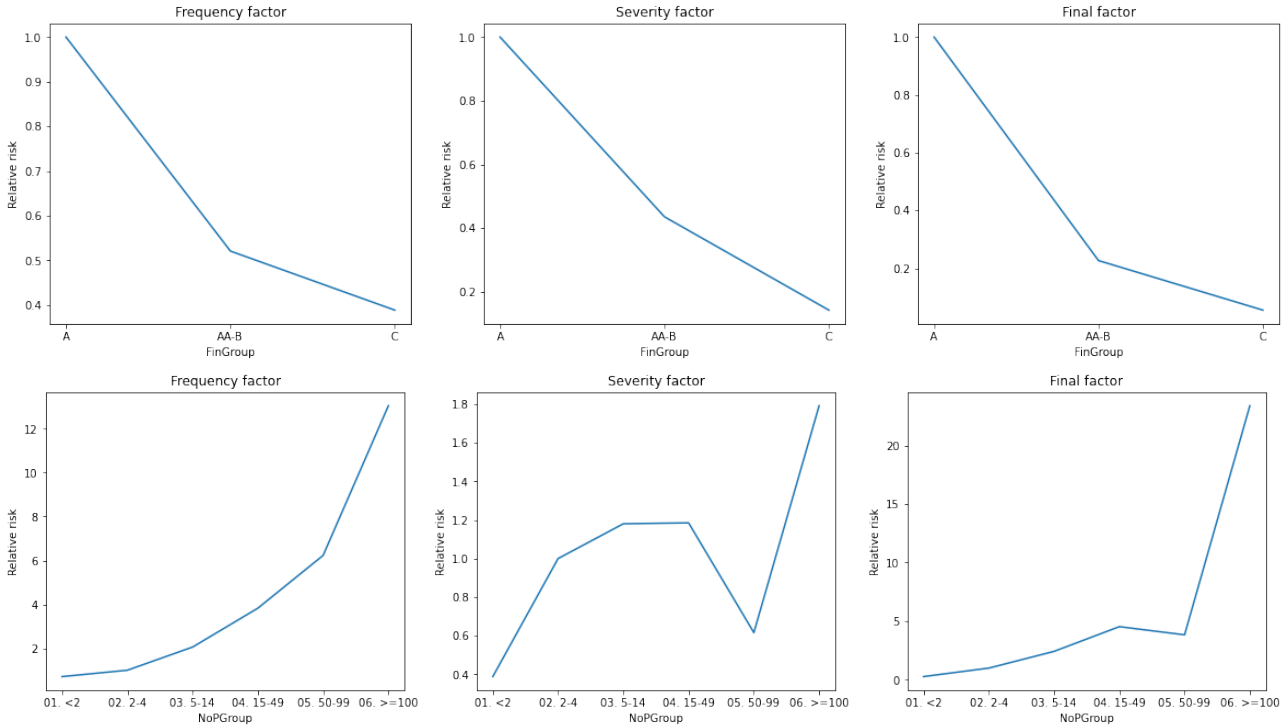
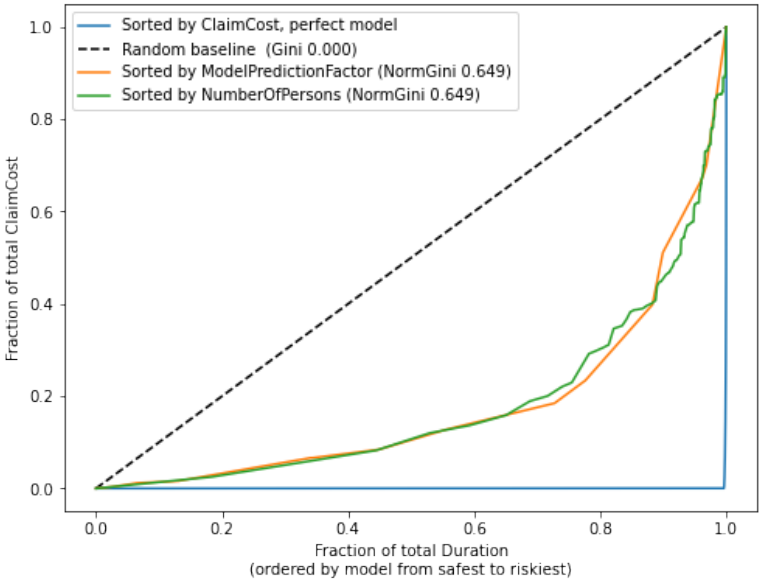


Figure 2: Three simple graphs

4.3 Combine models



4.4 Giniscore



4.5 Risk ratio validation

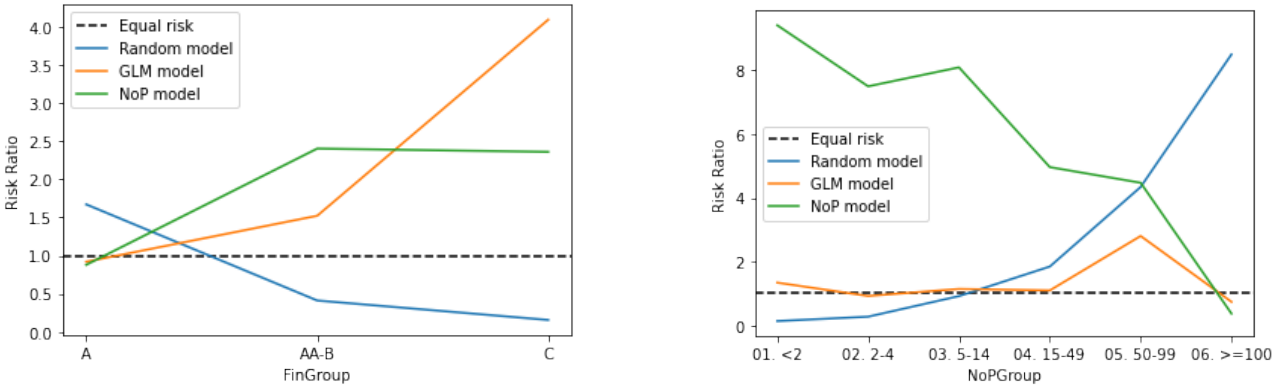


Figure 3: Risk ratio validation

4.6 Risk factors

Base level: 46.69998459348662

	FinGroup	NoPGroup	ModelPredictionFactor
0	1. AAA	01. < 2	0.275797
1	1. AAA	02. $2 - 4$	1.000000
2	1. AAA	03. $5 - 14$	2.426655
3	1. AAA	04. $15 - 49$	4.536105
4	1. AAA	05. $50 - 99$	3.840305
5	1. AAA	06. ≥ 100	23.389133
6	2. AA-B	01. < 2	0.062487
7	2. AA-B	02. $2 - 4$	0.226568
8	2. AA-B	03. $5 - 14$	0.549802
9	2. AA-B	04. $15 - 49$	1.027736
10	2. AA-B	05. $50 - 99$	0.870090
11	2. AA-B	06. $>= 100$	5.299229
12	3. C	01. < 2	0.015178
13	3. C	02. $2 - 4$	0.055034
14	3. C	03. $5 - 14$	0.133549
15	3. C	04. $15 - 49$	0.249642
16	3. C	05. $50 - 99$	0.211349
17	3. C	06. $>= 100$	1.287207

Table 5: Risk factors