Coding challenge

HUK-Coburg

26.11.2024

Bontempo, Federico

Die Challenge

Die Modellierung der zu erwartenden Schadenhöhe

Angenommen, ein Datensatz enthält Risikomerkmale und Informationen zu Ansprüchen. Das Ziel besteht darin, die erwartete Schadenhöhe pro Versicherungsnehmer und Jahr auf der Grundlage der Risikomerkmale des Kunden zu modellieren.

Herangehensweise

1. Clean dataset

2. Variablen untersuchen

3. Modell fitten

Variablenanpassung

Korrelationen mit der Zielvariable

Das passendste Modell finden

1. Clean dataset

How to clean

Das “string of a string” Format:

- Area
- VehBrand
- Region

(e.g. “ ‘A’ “)

String to integer:

- “B12” -> 12
- “B” -> 1

Addiere ClaimAmount = 0.0 für alle ClaimNb = 0

Mergen der 2 Datensets (nicht alle keys matchen, nicht alle Daten können verwendet werden)

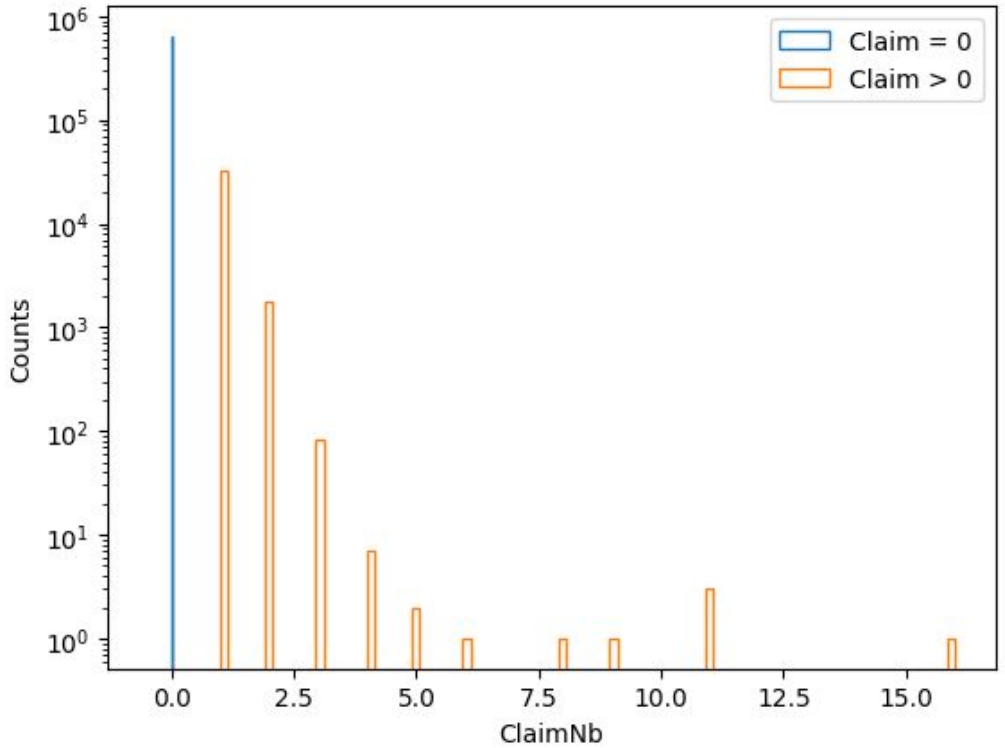
2. Variablen untersuchen

Variablen plotten

zB

ClaimNb:

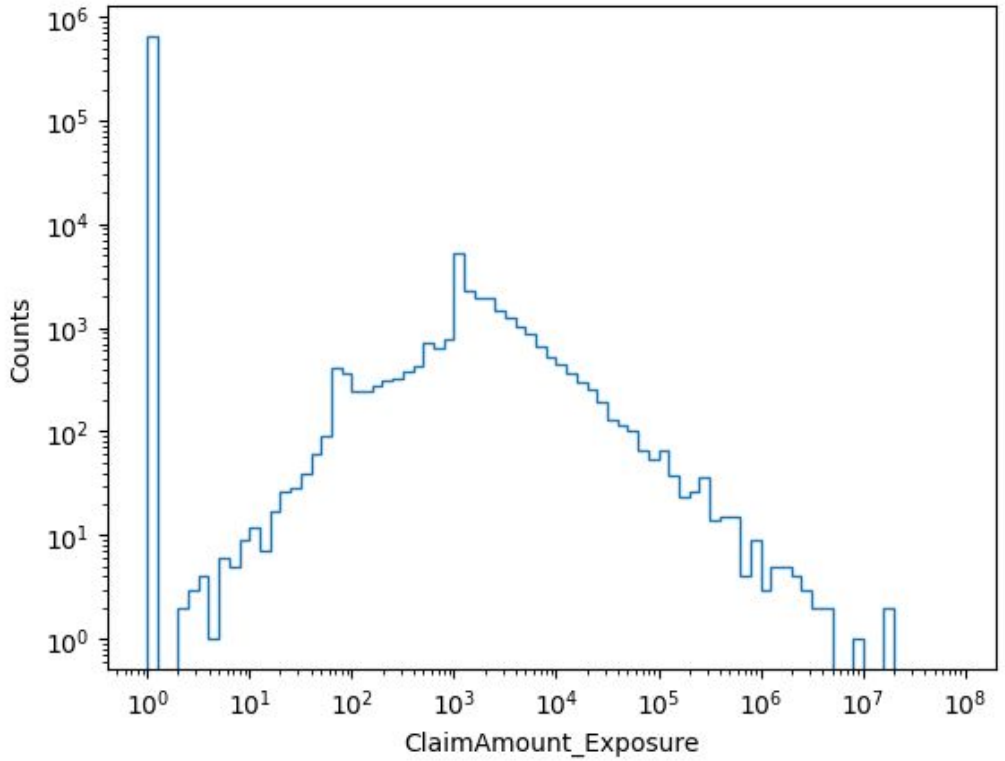
- Lineares Verhalten in semi-log
- Exponentieller Abfall



Die Zielvariable

ClaimAmount (+1) / Exposure:

- Ohne ClaimAmount = 0, sieht die Distribution gaußförmig in loglog Skala

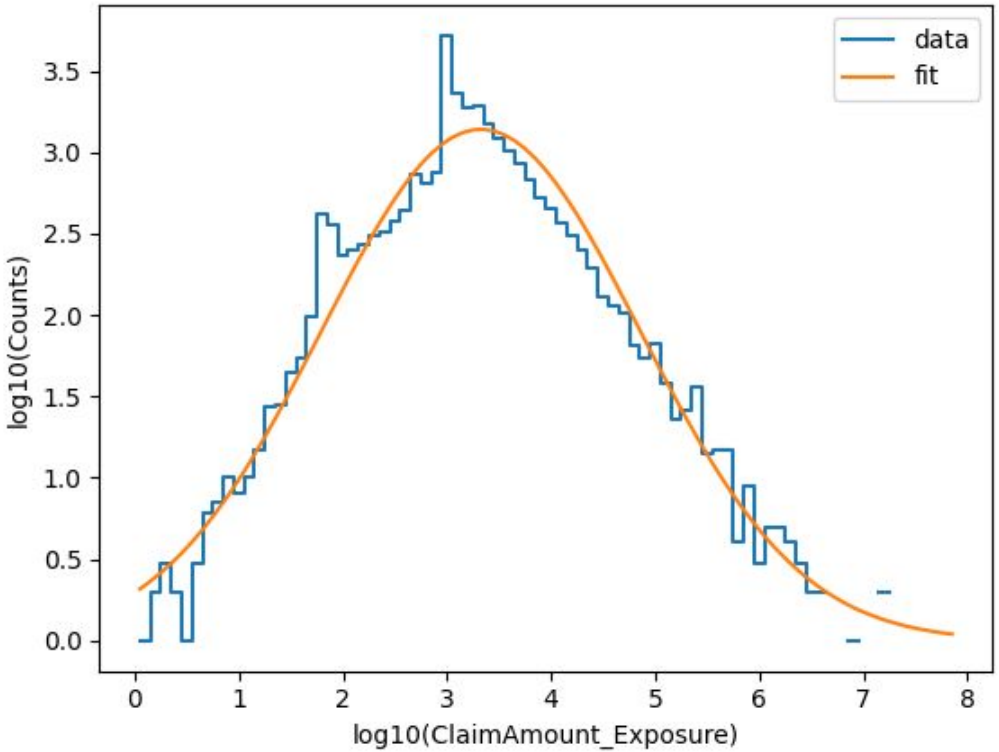


Die Zielvariable

ClaimAmount / Exposure:

- Ohne ClaimAmount = 0, sieht die Distribution gaußförmig in loglog Skala
- Gauß fitten

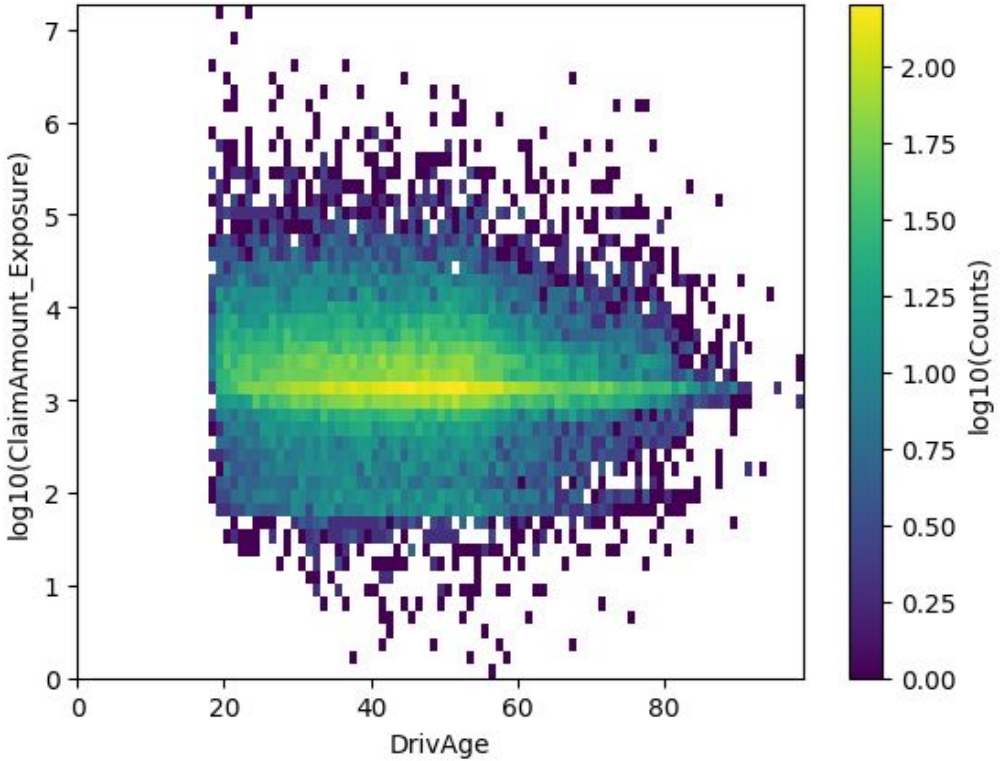
So kann man ClaimAmount / Exposure modellieren



Die Zielvariable

ClaimAmount / Exposure:
als Funktion von unabhängigen Variablen:

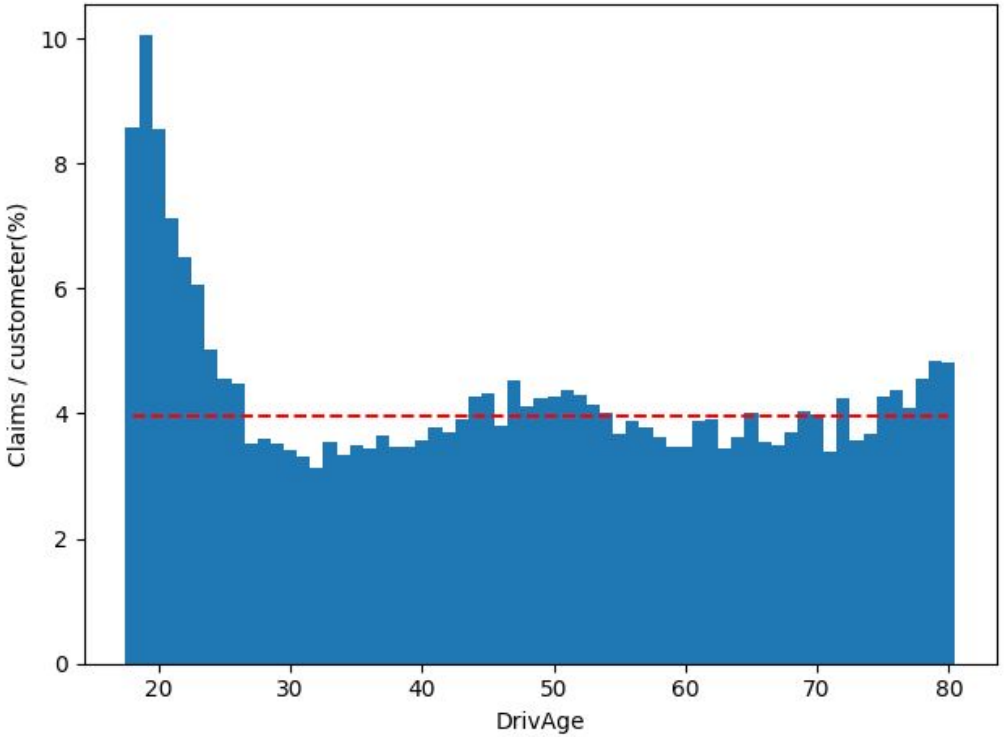
- gaußförmige Verteilung über alle Variablen und innerhalb der Variablen



Wahrscheinlichkeit des Claims

Untersuchen der Wahrscheinlichkeit das ein Claim auftreten kann als Funktion von allen Variablen:

- Im Vergleich zum Mittelwert: je größer die Fluktuation, desto signifikanter



3. Ein Modell finden und fitten

Das Modell: Generalized Linear Model

Linear

Einfaches lineares Modell zur Herstellung einer Beziehung zwischen den unabhängigen Variablen und der Zielvariable ClaimAmount / Exposure

Poisson

Kann zur Modellierung des diskreten ClaimNb verwendet werden

Tweedie

Die Potenz 1<p<2 kann zur Modellierung des ClaimAmount/Exposure mit einer Funktion zwischen einer Poisson- und einer Gammafunktion verwendet werden

Anmerkung: **statsmodels** und **sklearn** wurden für die Analyse in **python** verwendet

Linear

OLS: Ordinary least squares
mit DrivAge, BonusMalus als Variablen

Target: ClaimAmount/Exposure

OLS Regression Results

Dep. Variable:	y	R-squared:	0.000
Model:	OLS	Adj. R-squared:	0.000
Method:	Least Squares	F-statistic:	6.413
Date:	Mon, 25 Nov 2024	Prob (F-statistic):	0.00164
Time:	11:25:44	Log-Likelihood:	-6.4388e+06
No. Observations:	535116	AIC:	1.288e+07
Df Residuals:	535113	BIC:	1.288e+07
Df Model:	2		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-553.8950	259.505	-2.134	0.033	-1062.515	-45.275
DrivAge	-6.4852	4.101	-1.581	0.114	-14.524	1.554
BonusMalus	20.9456	5.997	3.493	0.000	9.192	32.699

Omnibus:	3113044.539	Durbin-Watson:	2.000
Prob(Omnibus):	0.000	Jarque-Bera (JB):	586961395734279.500
Skew:	376.178	Prob(JB):	0.00
Kurtosis:	162251.897	Cond. No.	531.

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Poisson

Poisson: kann man für diskrete Zahlen verwenden so wie ClaimNb.

Target: ClaimNb

Anschließend könnte man dann den/die ClaimAmount / Exposure schätzen, über die Modulation mittels der Gaußschen

Generalized Linear Model Regression Results

Dep. Variable:	ClaimNb	No. Observations:	535116
Model:	GLM	Df Residuals:	535106
Model Family:	Poisson	Df Model:	9
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-88716.
Date:	Mon, 25 Nov 2024	Deviance:	1.3680e+05
Time:	11:25:43	Pearson chi2:	5.87e+05
No. Iterations:	7	Pseudo R-squ. (CS):	0.006270
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-5.1878	0.051	-101.102	0.000	-5.288	-5.087
DrivAge	0.0117	0.001	22.477	0.000	0.011	0.013
BonusMalus	0.0212	0.000	54.114	0.000	0.020	0.022
VehAge	-0.0046	0.001	-3.451	0.001	-0.007	-0.002
fVehBrand	-0.0409	0.002	-23.114	0.000	-0.044	-0.037
VehPower	0.0338	0.004	9.614	0.000	0.027	0.041
fVehGas	-0.1295	0.014	-9.170	0.000	-0.157	-0.102
fArea	0.0703	0.006	10.844	0.000	0.058	0.083
Density	-1.571e-06	2.15e-06	-0.731	0.465	-5.78e-06	2.64e-06
fRegion	0.0006	0.000	2.307	0.021	8.4e-05	0.001

Tweedie

Tweedie: je nach Potenz p kann man verschiedene Funktionen haben.
Gauss, Poisson, Gamma.
Mit 1<p<2 liegt die Funktion zwischen einer Poisson- und einer Gammafunktion

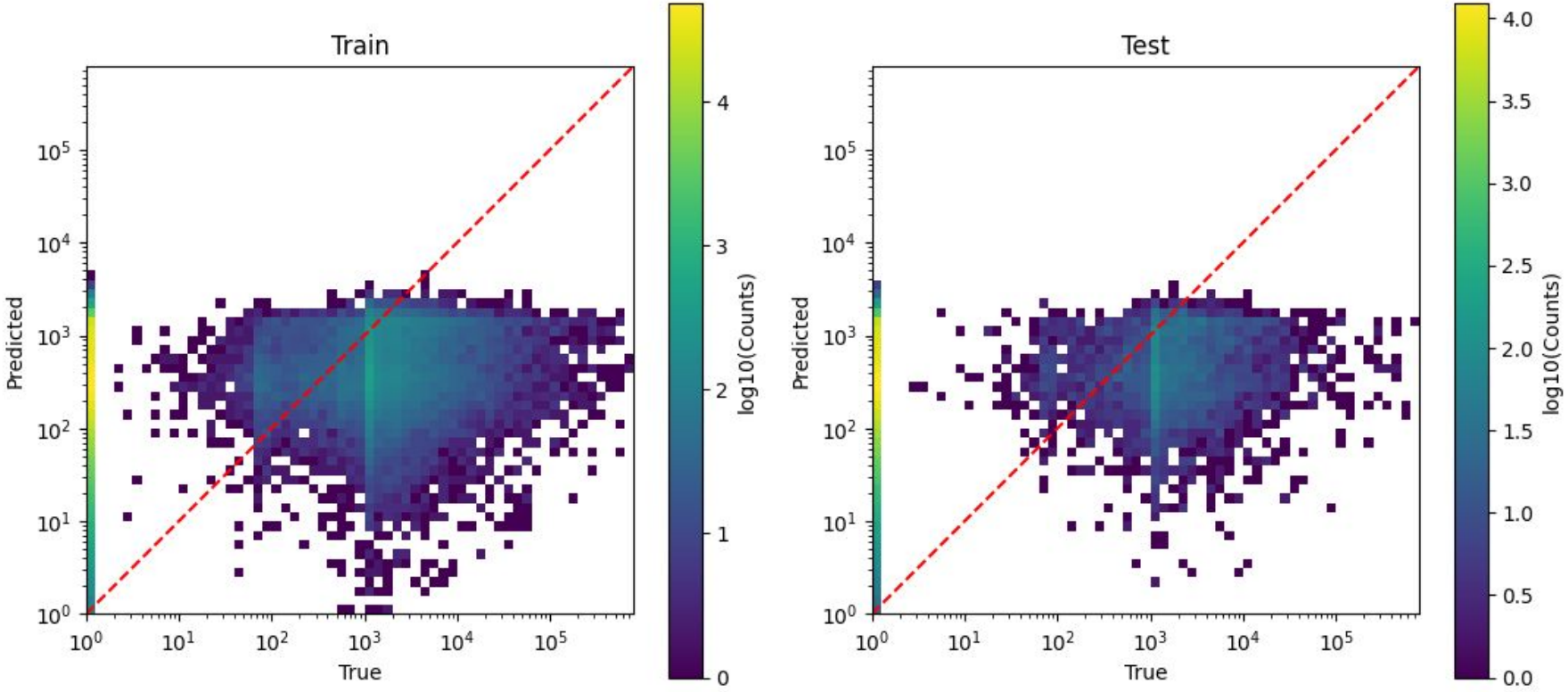
Target: ClaimAmount / Exposure

Generalized Linear Model Regression Results

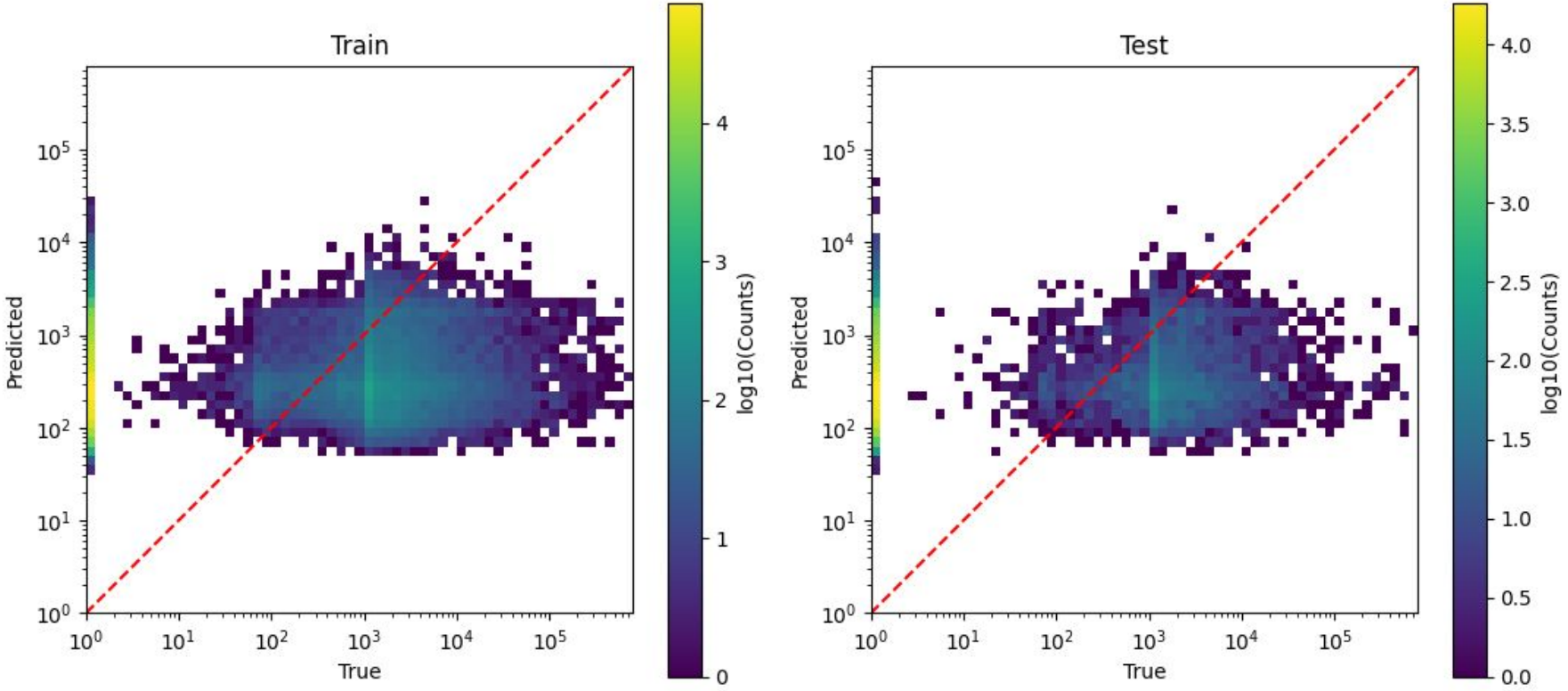
Dep. Variable:	ClaimAmount_Exposure	No. Observations:	535116
Model:	GLM	Df Residuals:	535113
Model Family:	Tweedie	Df Model:	2
Link Function:	Log	Scale:	1.7387e+06
Method:	IRLS	Log-Likelihood:	-780.82
Date:	Mon, 25 Nov 2024	Deviance:	2.7150e+09
Time:	11:25:54	Pearson chi2:	9.30e+11
No. Iterations:	14	Pseudo R-squ. (CS):	0.0001585
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	5.2110	0.566	9.202	0.000	4.101	6.321
DrivAge	-0.0263	0.008	-3.179	0.001	-0.042	-0.010
BonusMalus	0.0282	0.005	6.209	0.000	0.019	0.037

Vergleichsplots: OLS



Vergleichsplots: Tweedie



Zusammenfassung

Zusammenfassung / Outlook

Die Datensätze wurden für die Analyse bereinigt und zusammengeführt

Die Variablen wurden untersucht, um Korrelationen zwischen jeder Variable und dem Zielwert zu erkennen

Für die Analyse wurden verschiedene GLMs untersucht, hauptsächlich in zwei Versuchen:

1. Modellierung des/der ClaimAmounts/Exposure über eine Gauß-Verteilung und anschließend Abschätzung der Anzahl der Claims über eine Poisson-Anpassung
2. Modellierung des/der ClaimAmounts/Exposure über eine Tweedie-Verteilung.

Hinweis: Weitere Details im Code

Coding challenge

HUK-Coburg

26.11.2024

Bontempo, Federico

Links

https://www.statsmodels.org/stable/index.html

https://scikit-learn.org/stable/index.html

https://www.researchgate.net/publication/273578956_Auto_Insurance_Premium_Calculation_Using_Generalized_Linear_Models

https://www.researchgate.net/publication/369723727_Insurance_Risk_Prediction_Using_Machine_Learning

https://en.wikipedia.org/wiki/Generalized_linear_model