Credit Card Default Risk

Credit Card Default Risk

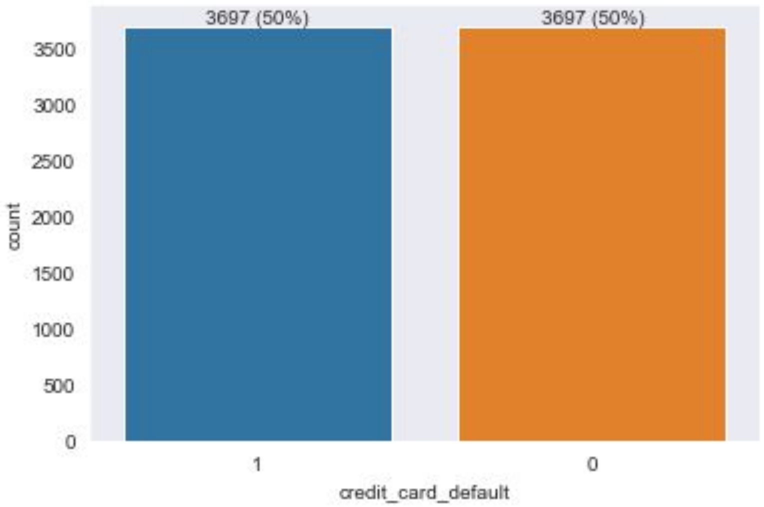
Task:
We are given relevant information about the company's customers. We're required to build a Machine Learning Model that can predict if there will be Credit Card Defaulters.

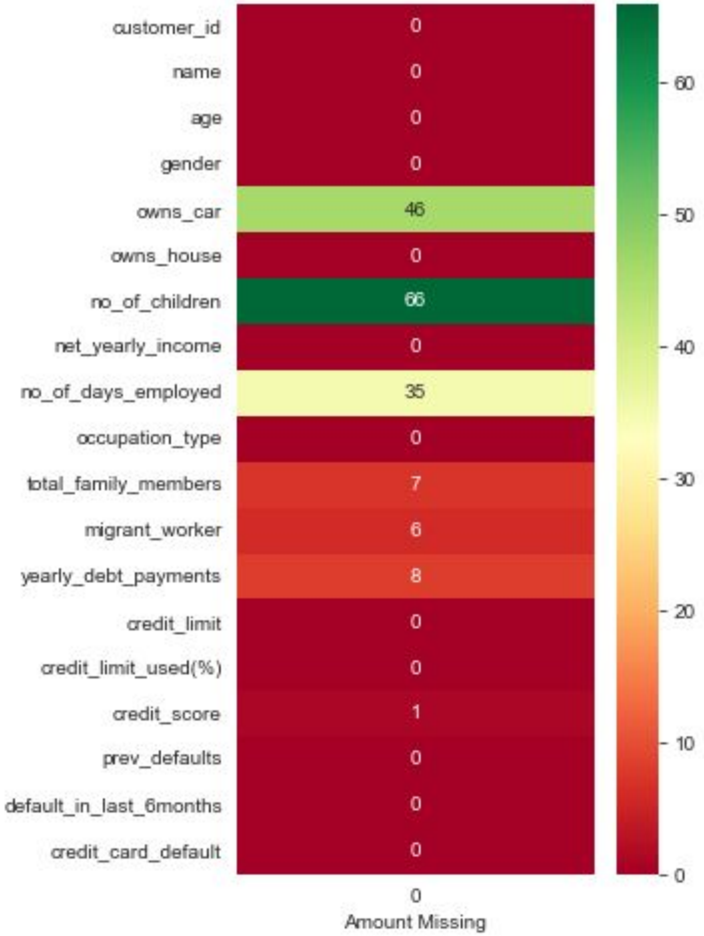
Dataset:
- Total samples: 7394
- Columns: 19

Column Name	Description
customer_id	unique identification of customer
name	name of customer
age	age of customer (Years)
gender	gender of customer (M or F)
owns_car	whether a customer owns a car (Y or N)
owns_house	whether a customer owns a house (Y or N)
no_of_children	number of children of a customer
net_yearly_income	net yearly income of a customer (USD)
no_of_days_employed	no. of days employed
occupation_type	occupation type of customer
total_family_members	no. of family members of customer
migrant_worker	customer is migrant worker (Yes or No)
yearly_debt_payments	yearly debt of customer (USD)
credit_limit	credit limit of customer (USD)
credit_limit_used(%)	credit limit used by customer
credit_score	credit score of customer
prev_defaults	no. of previous defaults
default_in_last_6months	whether a customer has defaulted (Yes or No)
credit_card_default	whether there will be credit card default (Yes or No)

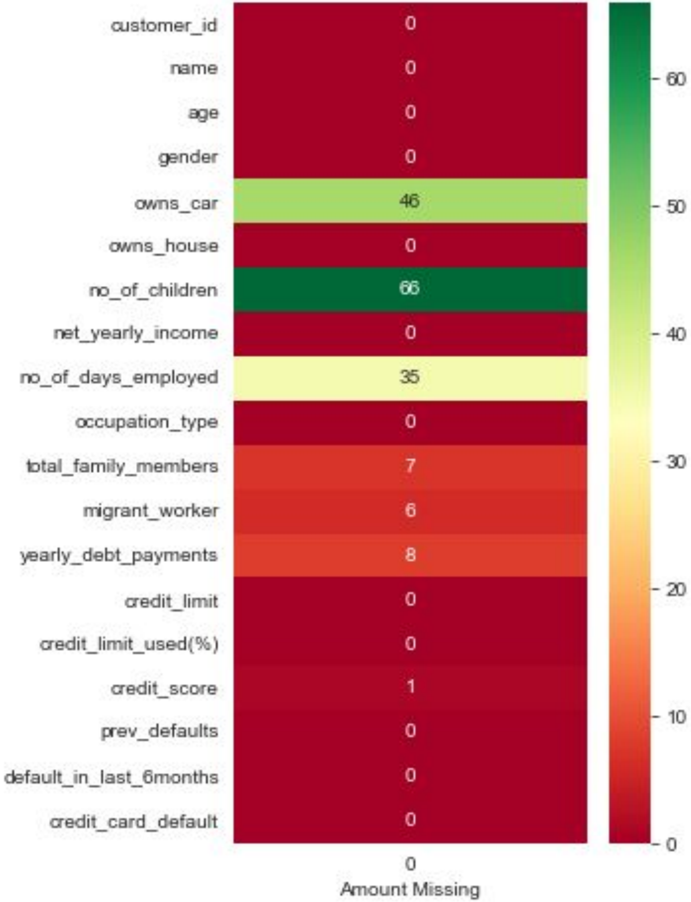
Credit Card Default Risk: description

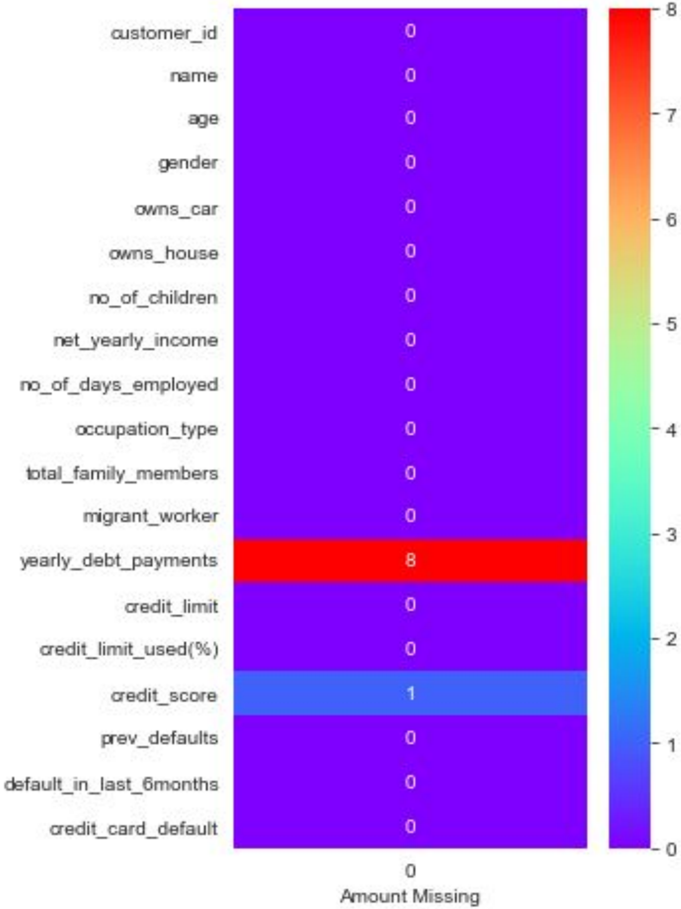
Dataset:
- Total samples: 7394
- Columns: 19





Filling in missing values: {'owns_car': 'N', 'no_of_children': 0, 'migrant_worker': 0, 'total_family_members': 1, 'no_of_days_employed': 0}

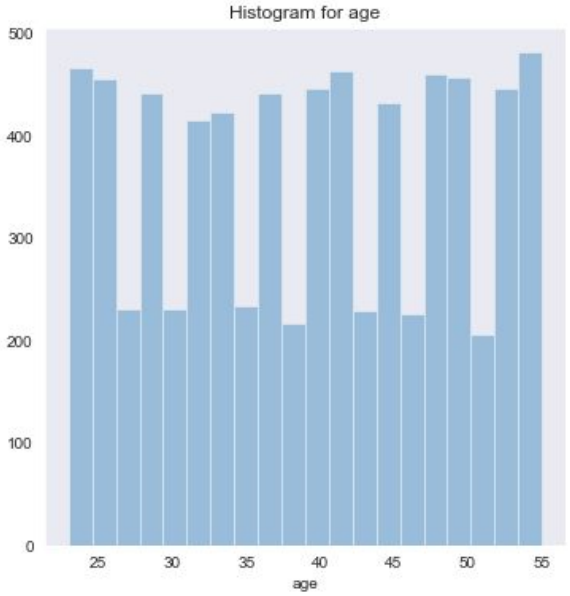


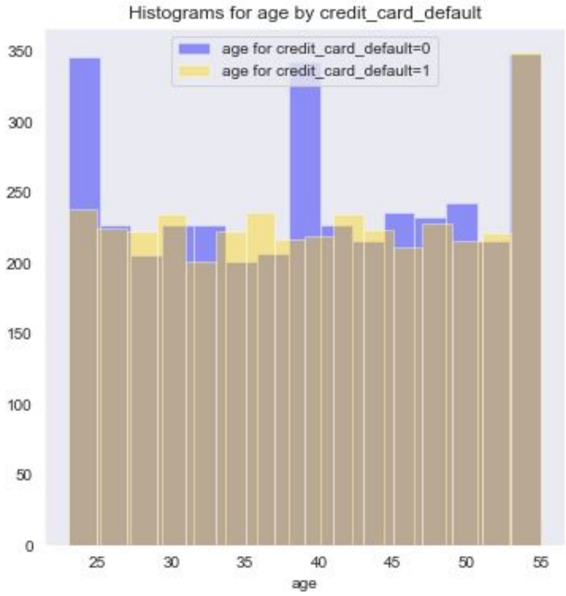


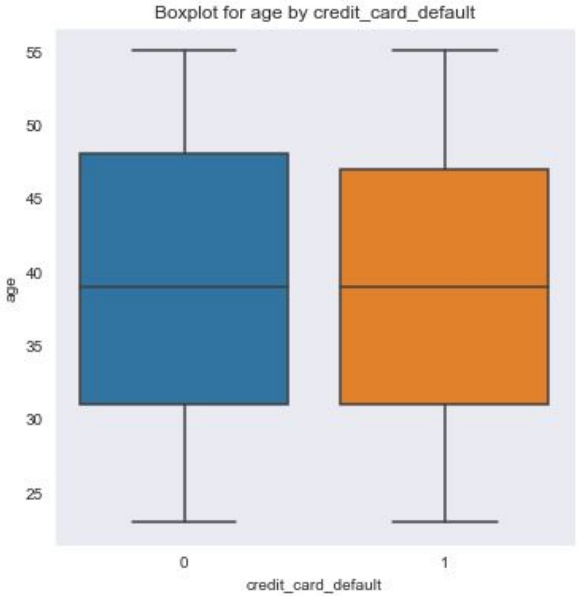
Credit Card Default Risk Dataset: Age

count	7394.000000	25%	31.000000
mean	39.053692	50%	39.000000
std	9.587768	75%	47.000000
min	23.000000	max	55.000000

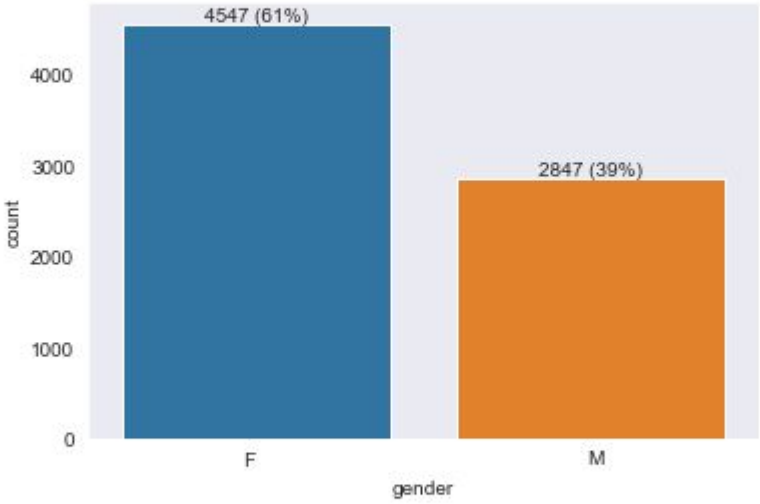
	age_split	credit_card_default
0	(22.999, 29.0]	0.511613
1	(29.0, 36.0]	0.500984
2	(36.0, 42.0]	0.500746
3	(42.0, 49.0]	0.496512
4	(49.0, 55.0]	0.488595



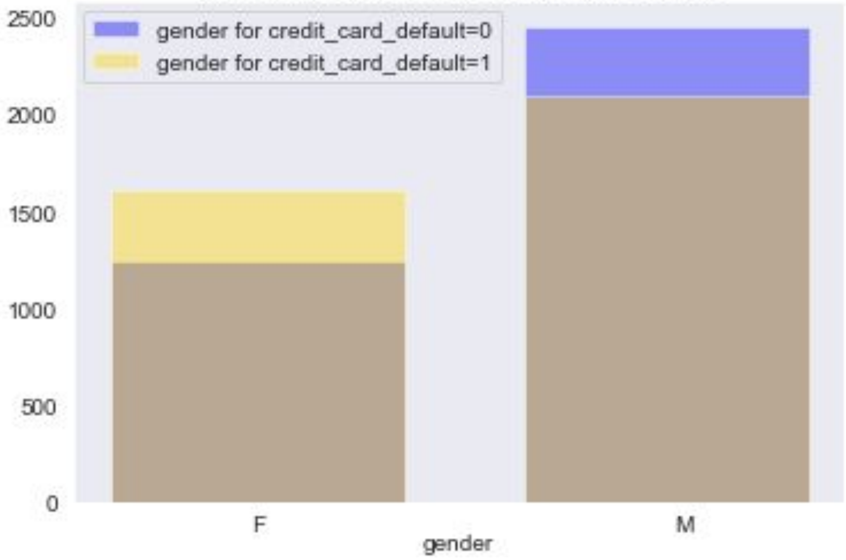




Credit Card Default Risk Dataset: gender



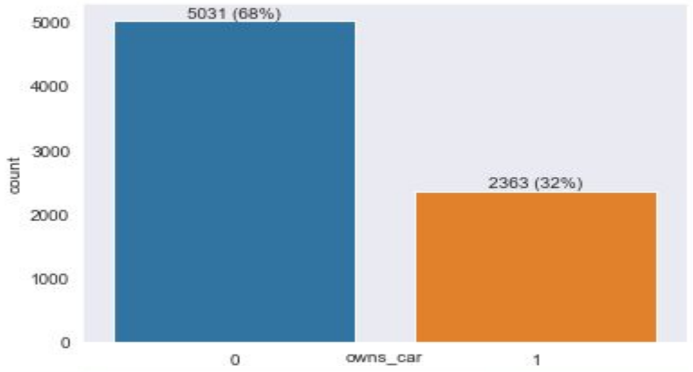
	gender	credit_card_default
1	M	0.563400
0	F	0.460303

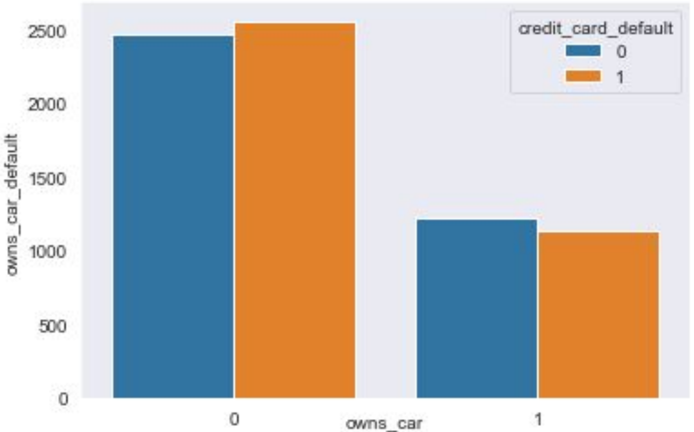


Credit Card Default Risk Dataset:

owns_car

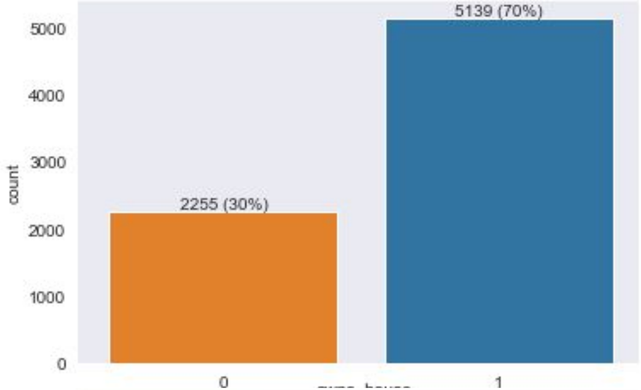
	owns_car	credit_card_default
0	0	0.508845
1	1	0.481168

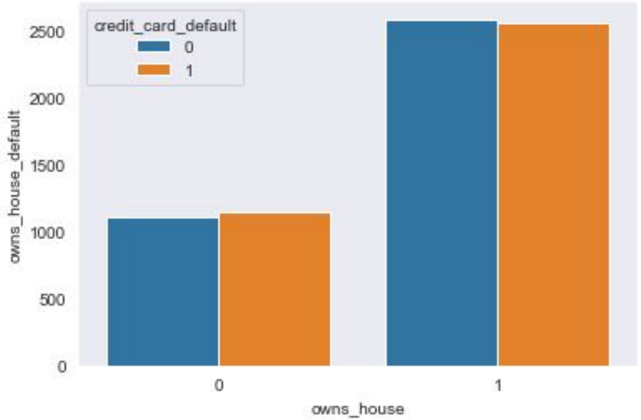




owns_house

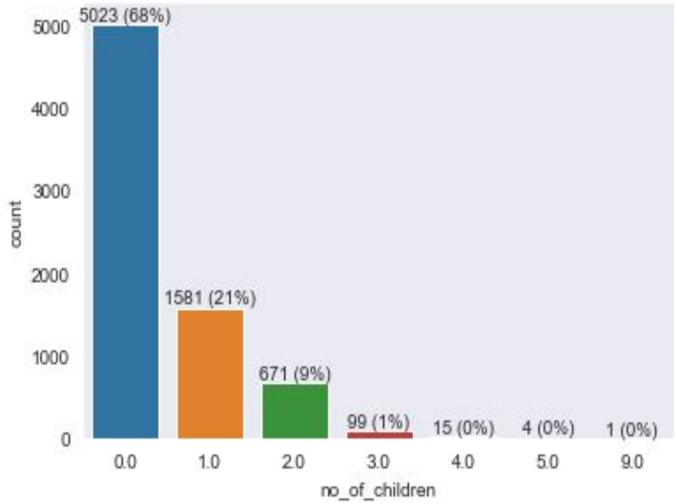
	owns_house	credit_card_default
0	0	0.506874
1	1	0.496984

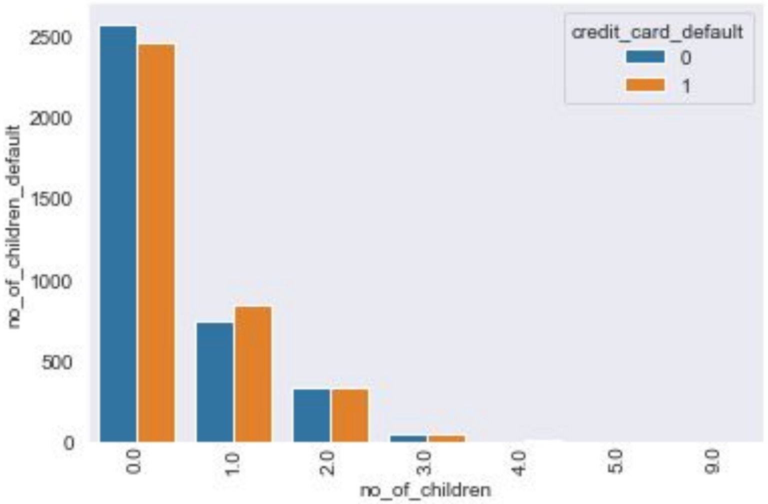




Credit Card Default Risk Dataset: <u>no_of_children</u>

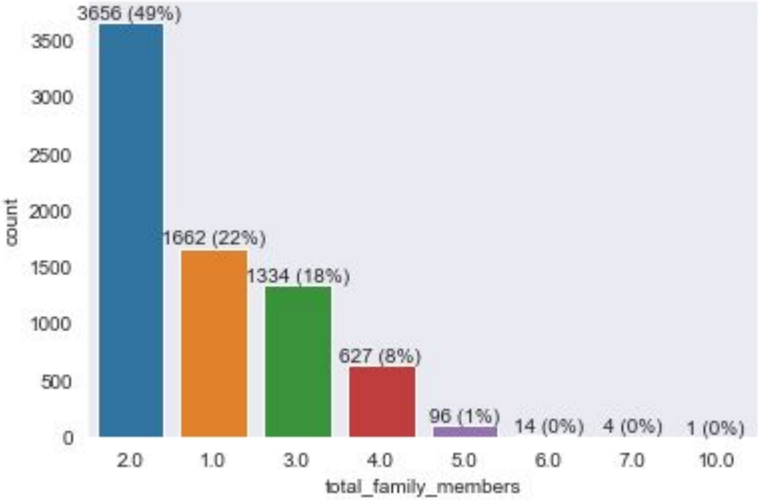
	no_of_children	credit_card_default
6	9.0	1.000000
5	5.0	0.750000
4	4.0	0.733333
1	1.0	0.530677
2	2.0	0.506706
3	3.0	0.505051
0	0.0	0.488354

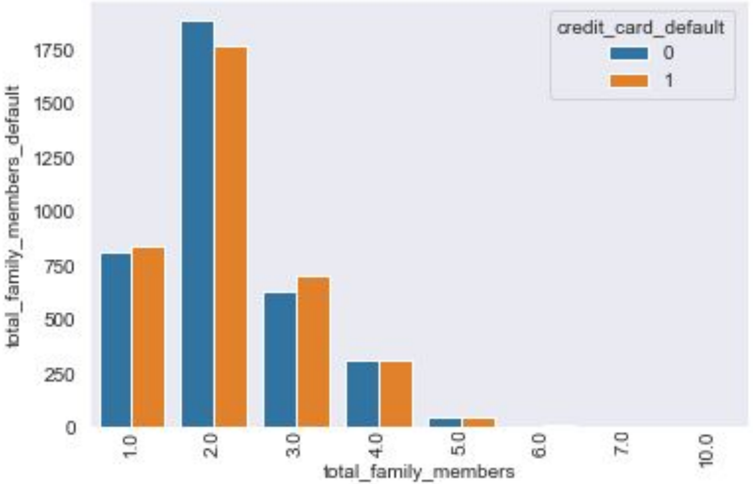




Credit Card Default Risk Dataset: total_family_members

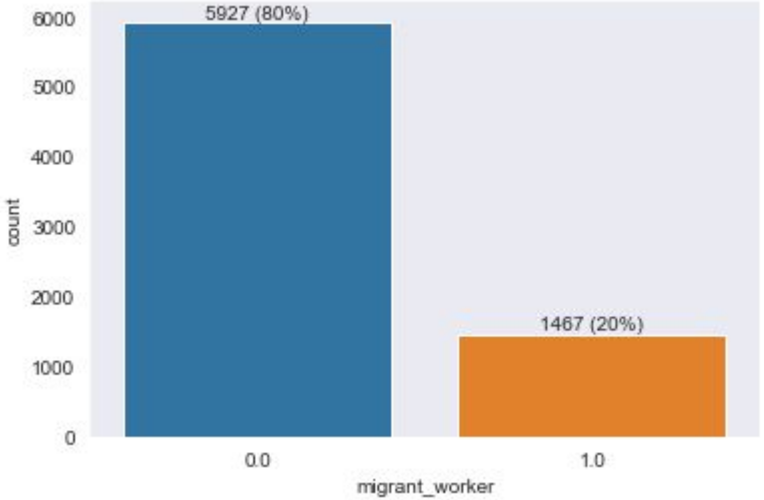
	total_family_members	credit_card_default
7	10.0	1.000000
6	7.0	0.750000
5	6.0	0.714286
2	3.0	0.526987
4	5.0	0.510417
0	1.0	0.508424
3	4.0	0.502392
1	2.0	0.484409

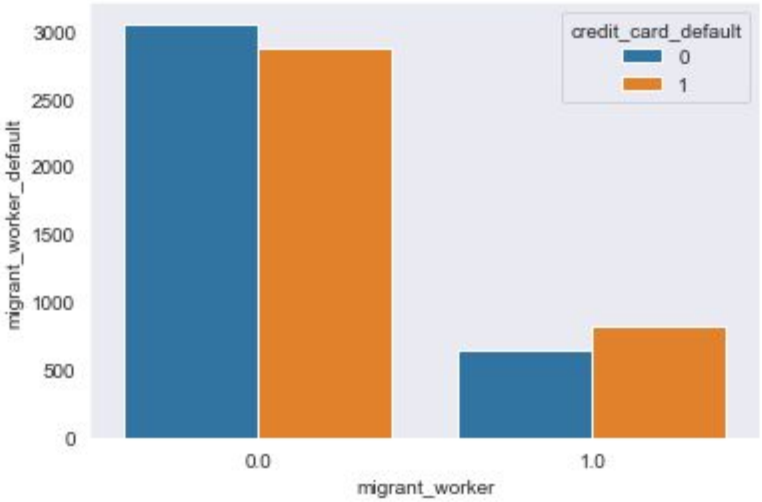




Credit Card Default Risk Dataset: <u>migrant_worker</u>

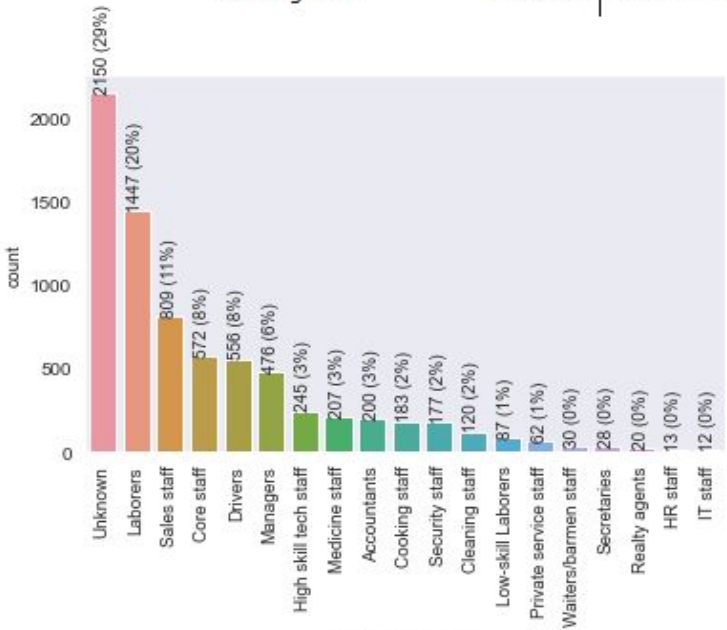
	migrant_worker	credit_card_default
1	1.0	0.561009
0	0.0	0.484900

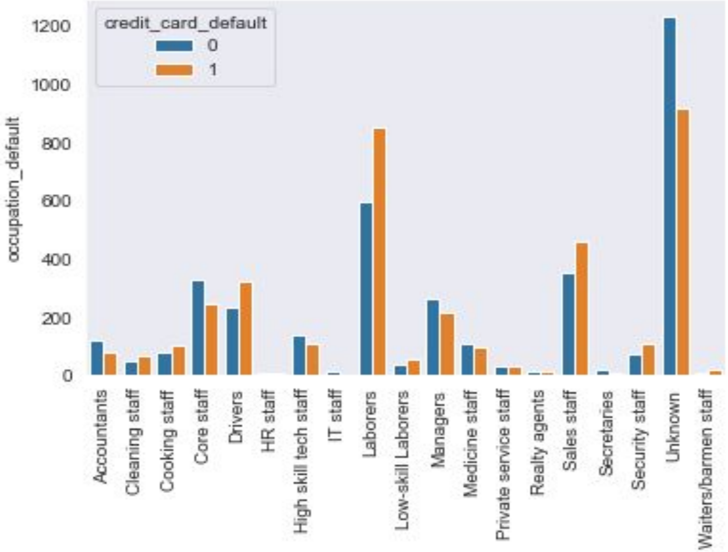




Credit Card Default Risk Dataset: occupation_type

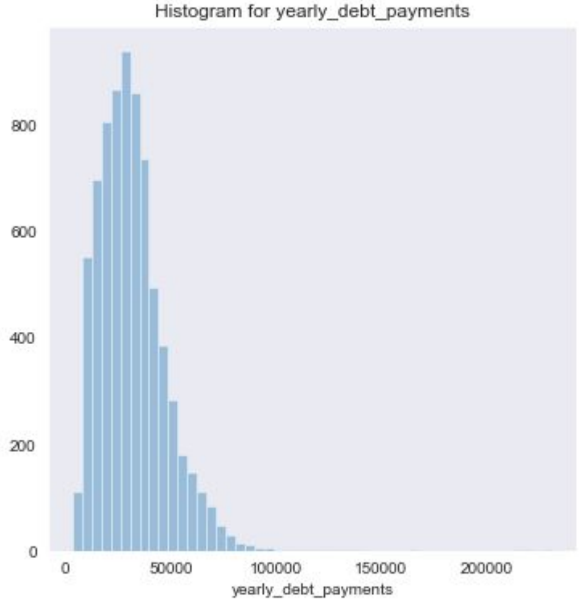
occupation_type	credit_card_default
Waiters/barmen staff	0.700000
Low-skill Laborers	0.609195
Security staff	0.598870
Laborers	0.588113
Drivers	0.577338
Cleaning staff	0.575000
Sales staff	0.566131
Cooking staff	0.562842
HR staff	0.538462
Realty agents	0.500000
Medicine staff	0.478261
Private service staff	0.467742
Managers	0.449580
High skill tech staff	0.436735
Unknown	0.426977
Core staff	0.424825
Accountants	0.390000
Secretaries	0.285714
IT staff	0.166667

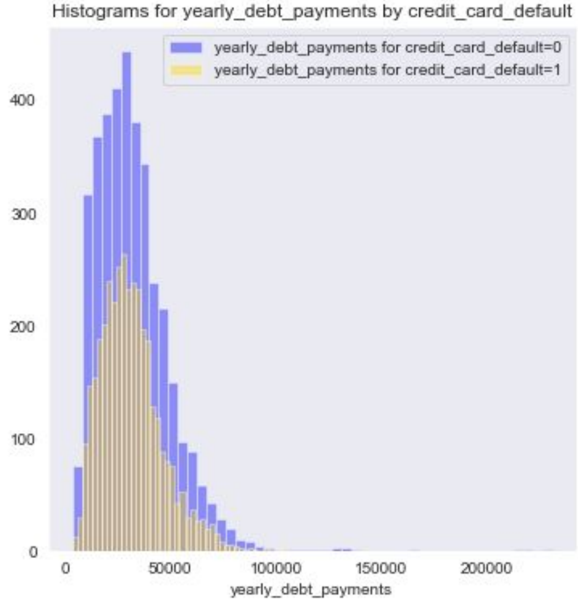


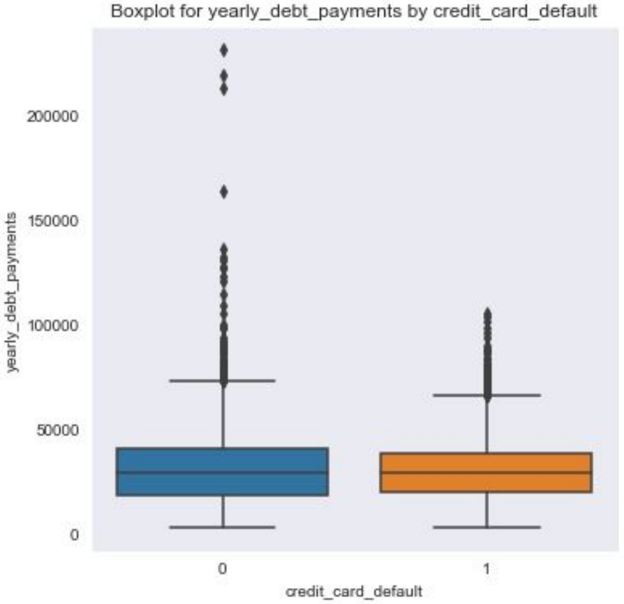


Credit Card Default Risk Dataset: yearly_debt_payments

	yearly_debt_payments		
count	7386.000000	25%	19739.872500
mean	31375.668179	50%	29137.605000
std	16229.100721	75%	39519.390000
min	3256.330000	max	231222.570000



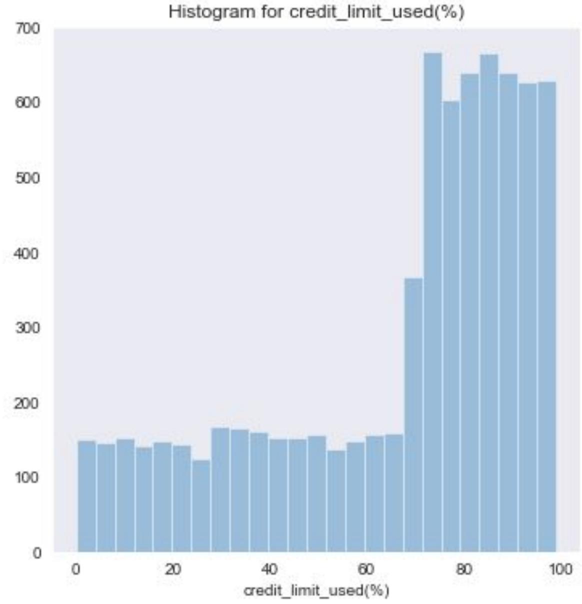


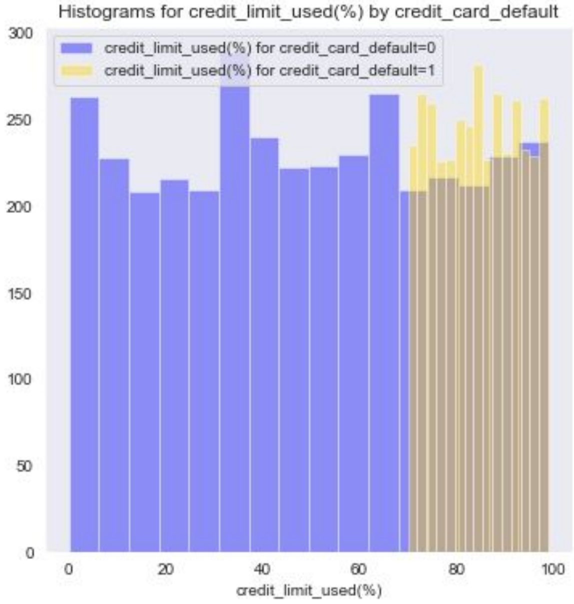


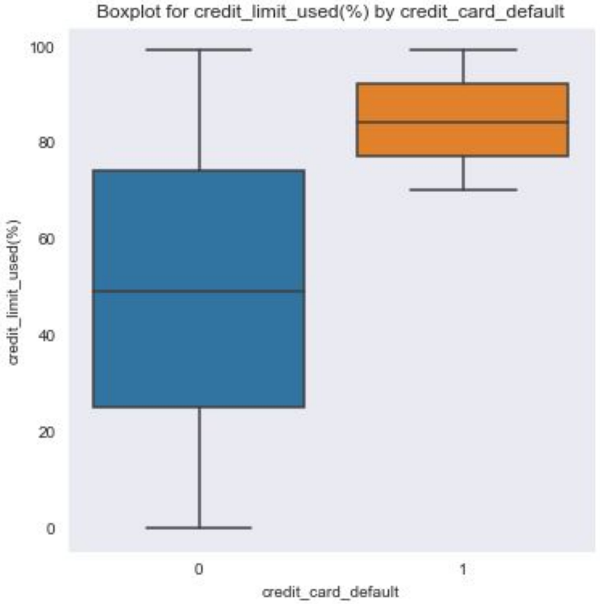
Credit Card Default Risk Dataset: credit_limit_used(%)

	credit_limit_used(%)_split	credit_card_default
0	(-0.001, 65.0]	0.000000
1	(65.0, 84.0]	0.733728
2	(84.0, 99.0]	0.772498

min(credit_limit_used(%) = 68

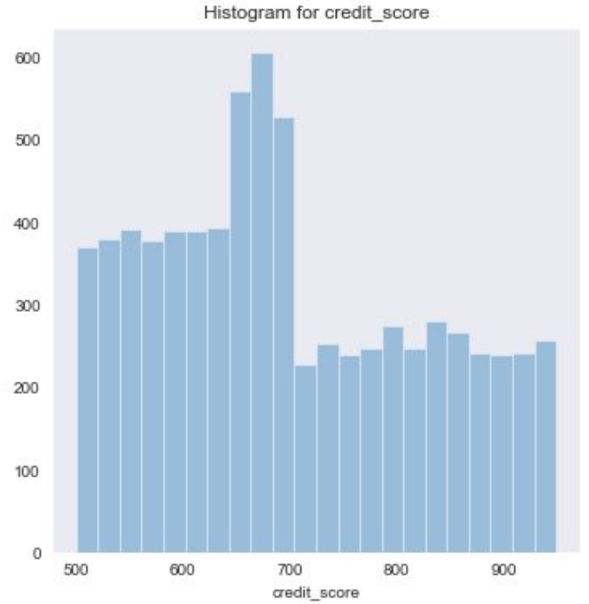


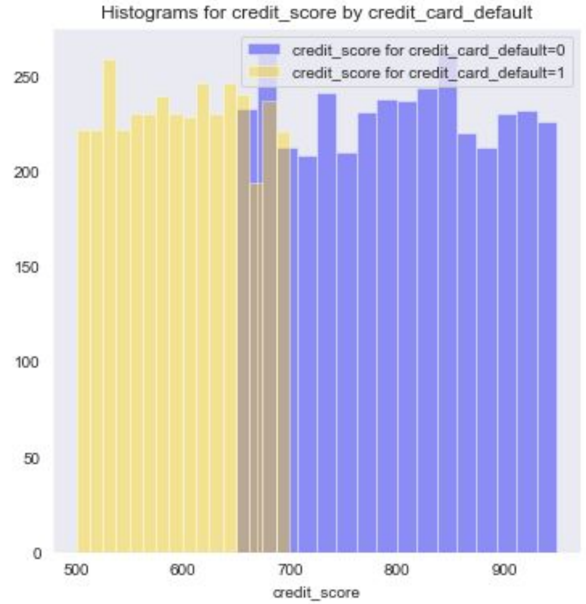


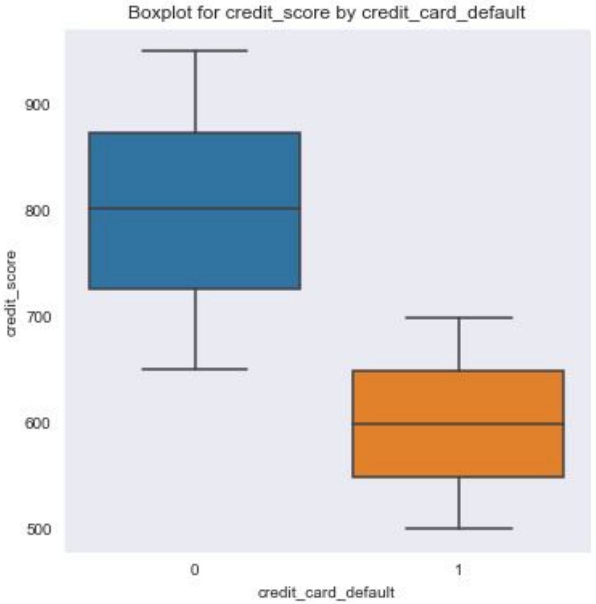


Credit Card Default Risk Dataset: credit_score

	credit_score_split	credit_card_default
0	(499.999, 580.0]	1.000000
1	(580.0, 654.0]	0.958700
2	(654.0, 709.0]	0.534247
3	(709.0, 830.0]	0.000000
4	(830.0, 949.0]	0.000000



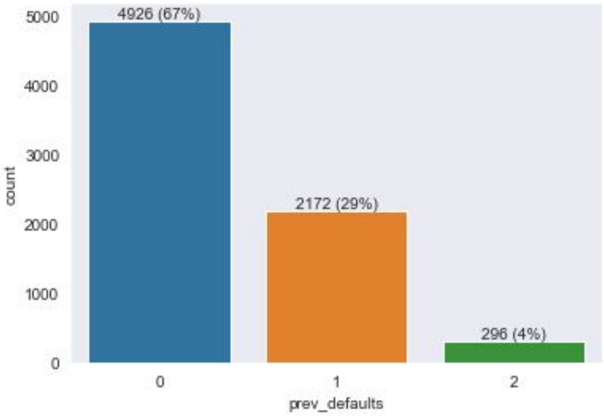


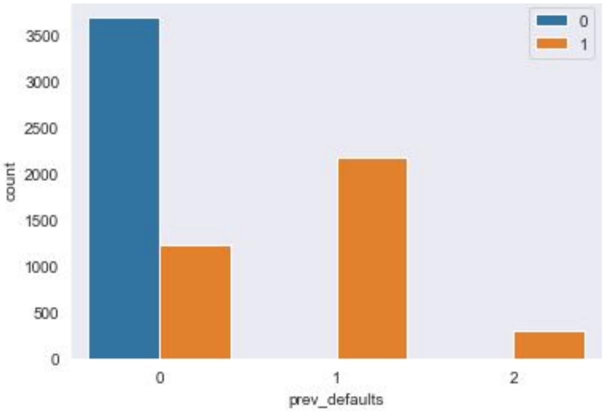


Credit Card Default Risk Dataset:

prev_defaults

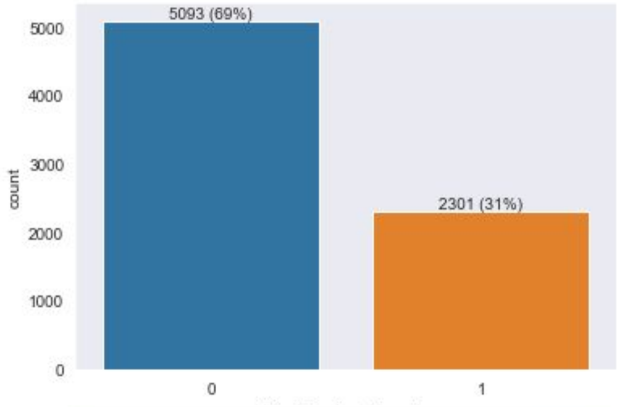
prev_defaults	credit_card_default
1	1.000000
2	1.000000
0	0.249492

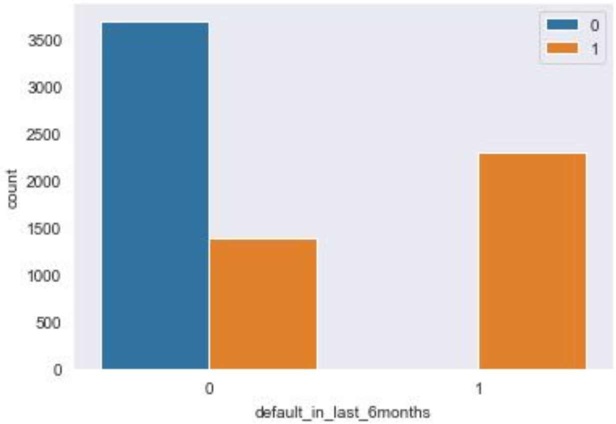




default_in_last_6months

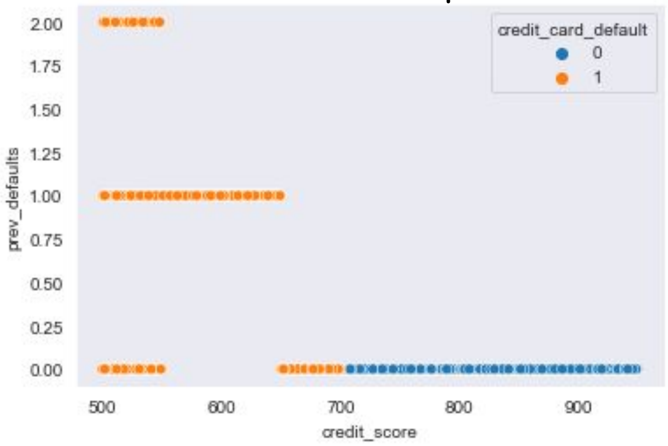
	credit_card_default
1	1.000000
0	0.274102



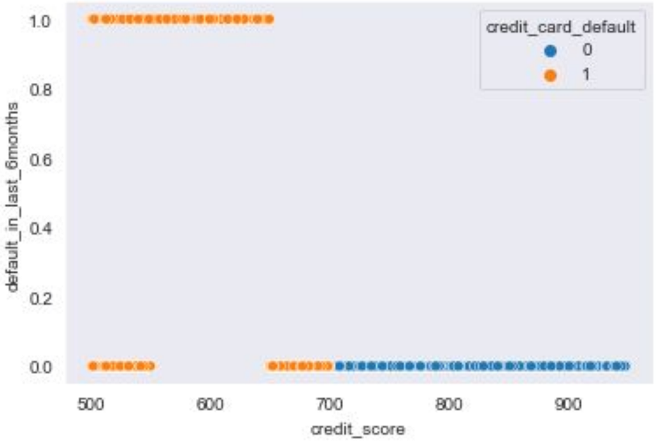


Credit Card Default Risk Dataset:

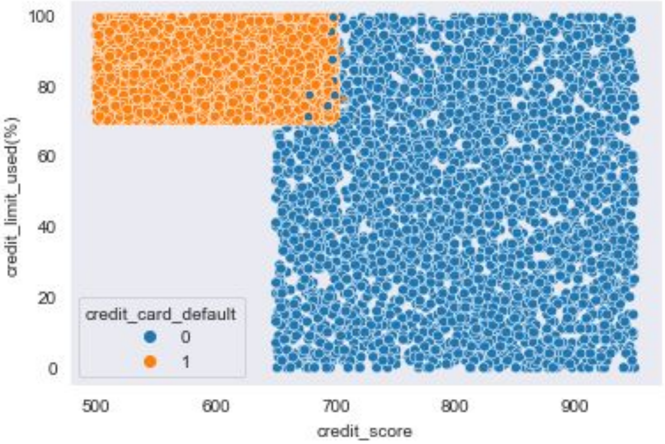
Credit card default on basis of 'credit score' and 'prev_defaults'



Credit card default on basis of 'credit score' and 'default_in_last_6months'



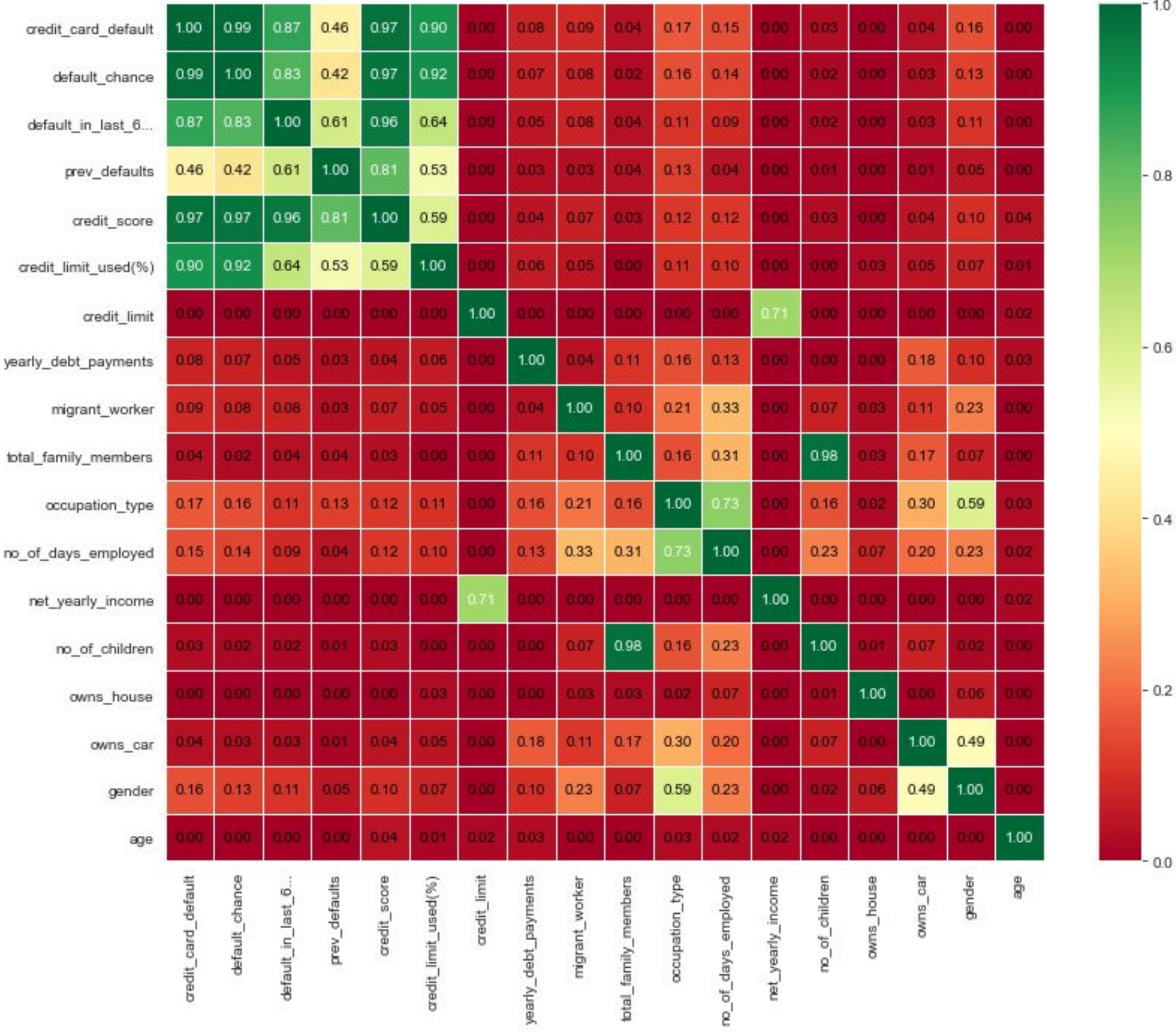
Credit card default on basis of 'credit score' and 'credit limit used'



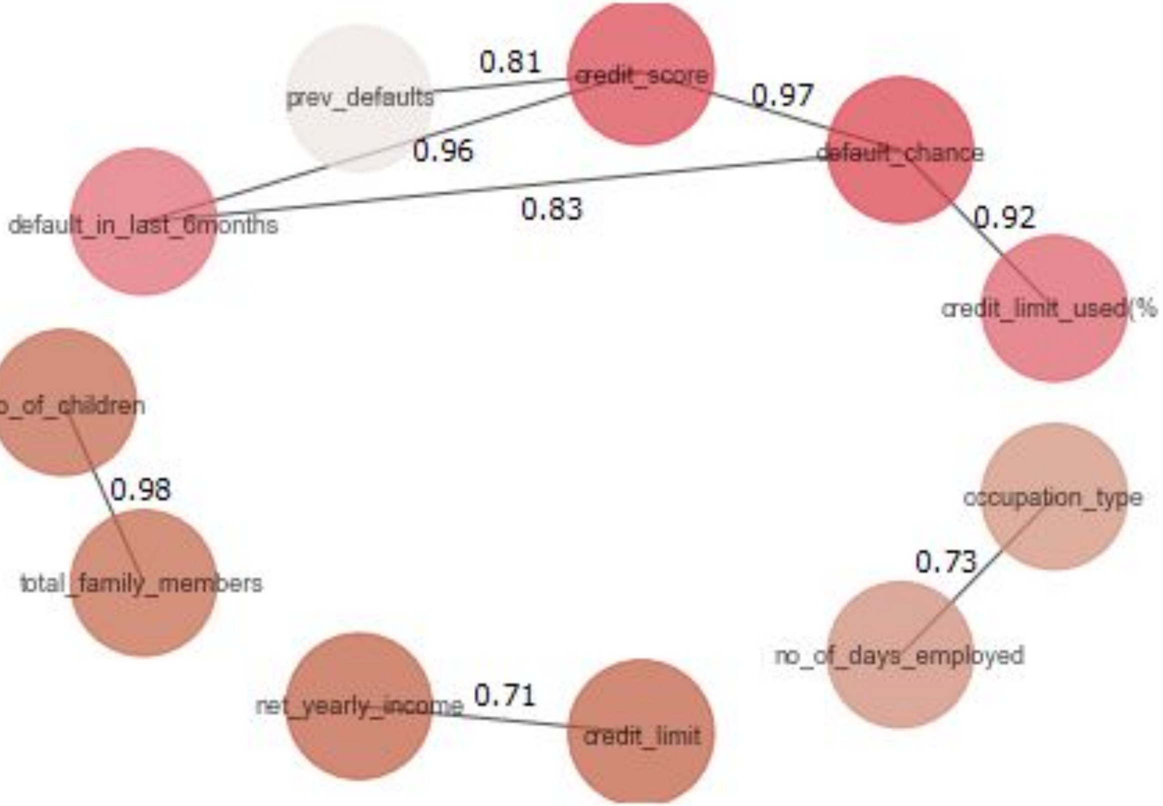
'credit_limit_used(%)' > 68
& **'credit_score'** < 700

1 3697
0 191

Correlogram

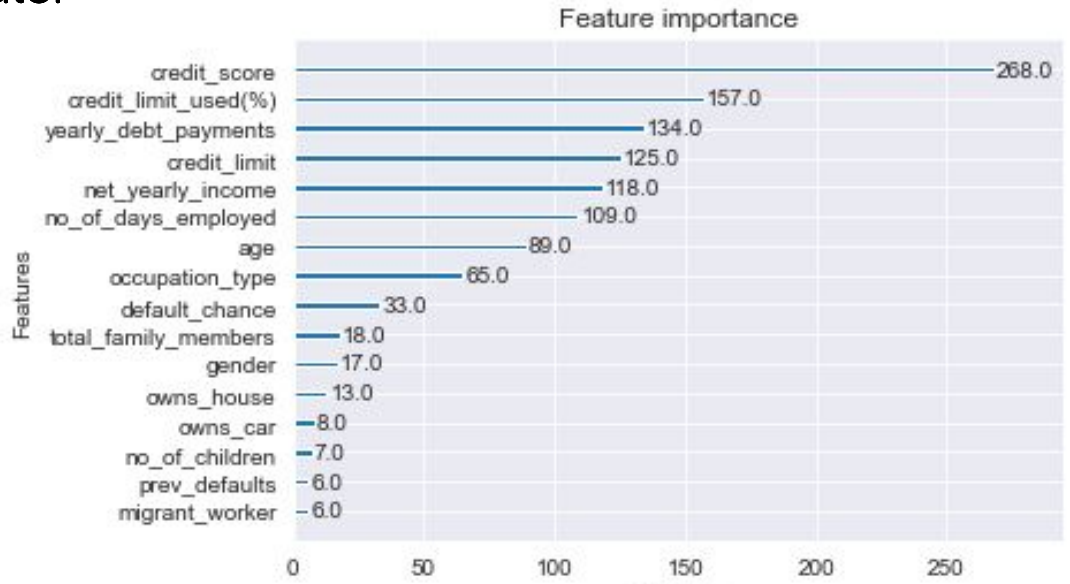


Correlation analysis as Graph



Preprocessing data:

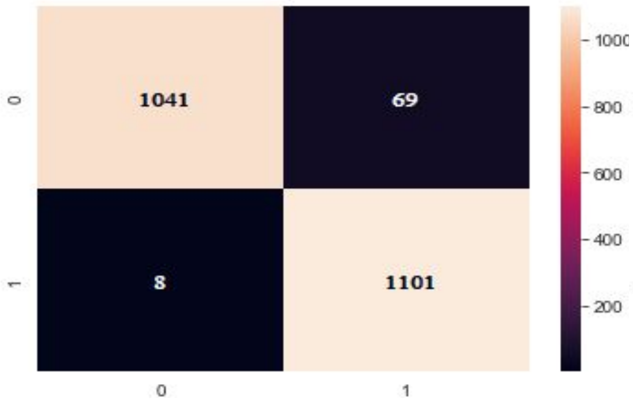
1. Filling in missing values
2. Creating feature where 'credit_limit_used(%)' > 68 & 'credit_score' < 700
3. Drop column 'customer_id', 'name'
4. Encoding labels for categorical features (gender, owns_car, owns_house, migrant_worker, occupation_type) with LabelEncoder
5. Scaling data with StandardScaler
6. SMOTE method for train_data
7. Feature Selection with GradientBoostingClassifier



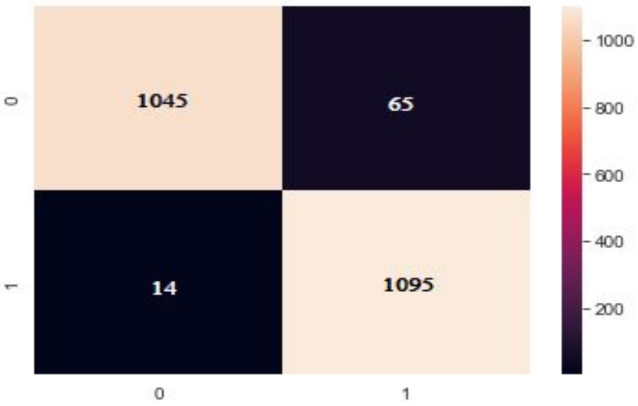
Model comparison

Algorithms	Features	Accuracy	
		train	test
Logistic Regression	13	97.4299	96.5299
K-Nearest Neighbours	13	97.2560	96.4398
Random Forest	2	97.7391	96.8004
NN	17	97.5348	96.6754

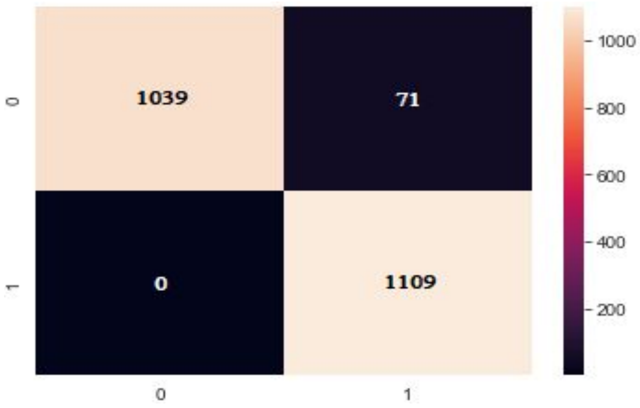
Logistic Regression



K-Nearest Neighbours

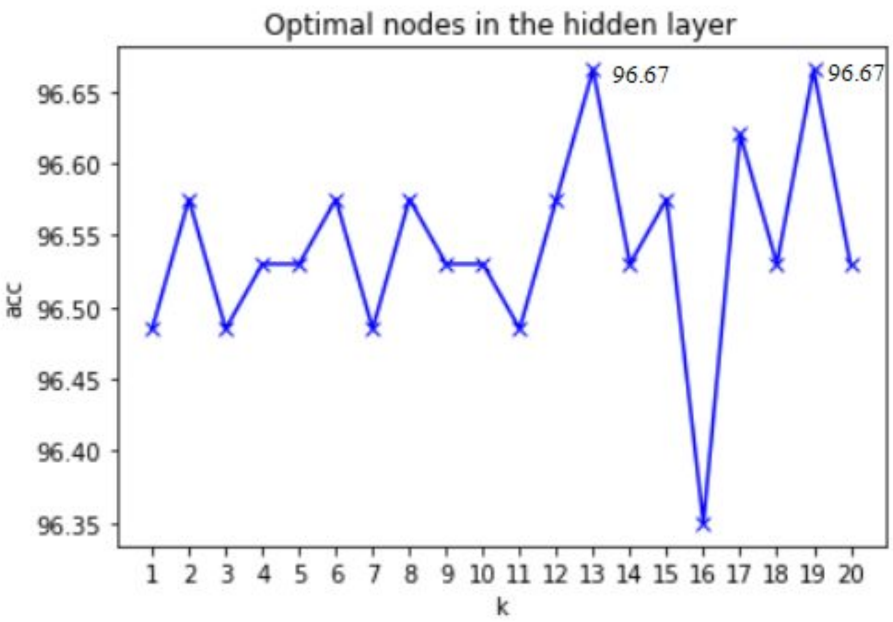


Random Forest



Model

- The model expects data with 17 variables
- The first hidden layer has from 1 to 20 nodes and uses the sigmoid activation function.
- The second hidden layer has 8 nodes and uses the relu activation function.
- The output layer has one node and uses the sigmoid activation function.
- Loss function binary_crossentropy
- Optimizer - adam
- Metrics - accuracy

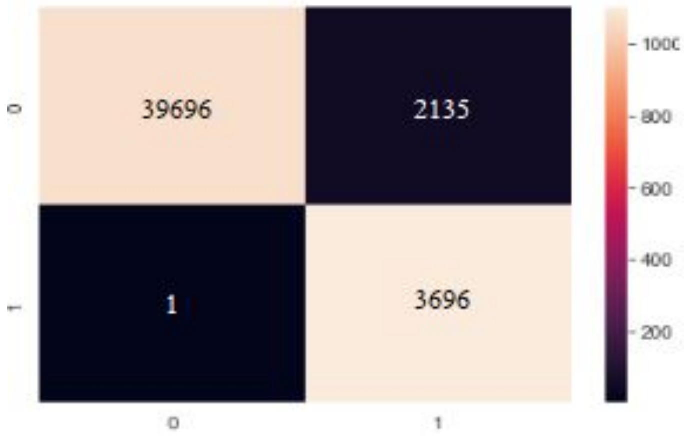


Random Forest

Dataset:

- Total samples: 45528
- Columns: 2

Algorithms	Features	Accuracy	
		train	test
Random Forest	2	97.7391	95.3084



Conclusion:

1. **Credit card default for clients with** 'credit_limit_used(%)' > 68 & 'credit_score' < 700
2. **Correlation between:**
   a. net_yearly_income - credit_limit (0.71)
   b. no_of_days_employed - occupation_type (0.73)
   c. no_of_children - total_family_members (0.98)
   d. prev_defaults - credit_score (0.81)
   e. default_in_last_6m - default_chance (0.83)
   f. credit_limit_used(%) - default_chance (0.92)
   g. default_in_last_6m - credit_score (0.96)
   h. credit_score - default_chance (0.97)
3. **Important features:** credit_score; credit_limit_used(%); yearly_debt_payments; credit_limit; net_yearly_income; no_of_days_employed.
4. **The highest accuracy :**

   Random Forest: train = 97.7391 ; test = 96.8004.
5. **Accuracy for all data: 95.3084**