Teo Hui Jie

Credit Card Default Prediction Model Analysis Report

Introduction and Problem Statement

The purpose of this report is to develop a predictive model that evaluates the credit risk of customers by forecasting the likelihood of default on their credit card payments within the next 12 months. This will help the bank gain deeper insights into the risk profile of its new credit card portfolio and support informed decisions regarding potential adjustments to its credit policies or debt recovery strategies.

Feature Engineering

To improve the predictive power of the model, three new features were created from the existing dataset. The first variable is **Worst Payment Status** which highlights the worst (highest numeric) payment status in the past three billing cycles. The second variable is **Continuous Delinquency** which is a binary variable indicating whether a customer has delayed payments ($payment\ status \geq 2$) in all three months (June, July and August). The last variable is **Average Payment Ratio** which measures the ratio of total payments made to total due amounts.

Data Preparation

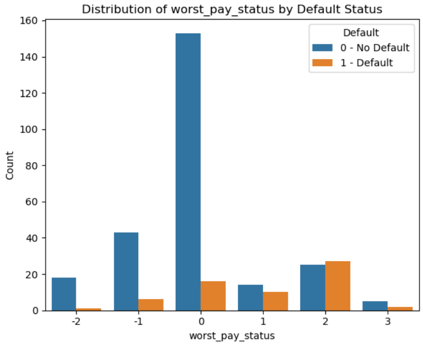
First, the dataset was assessed for missing values, and none were identified. The “id” column was removed as it contained entirely unique entries and did not hold any predictive relationship with “default”. Furthermore, categorical variables including “education” “marriage” and “owned_rent” were transformed through one-hot encoding to convert them into numerical format, allowing for more effective analysis and model development.

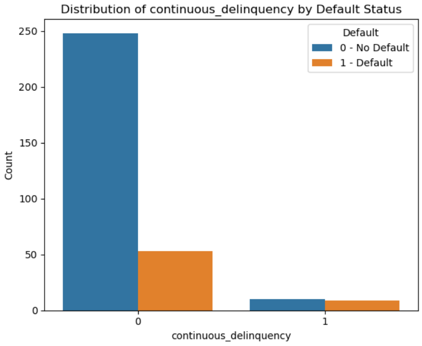
Exploratory Data Analysis (EDA)

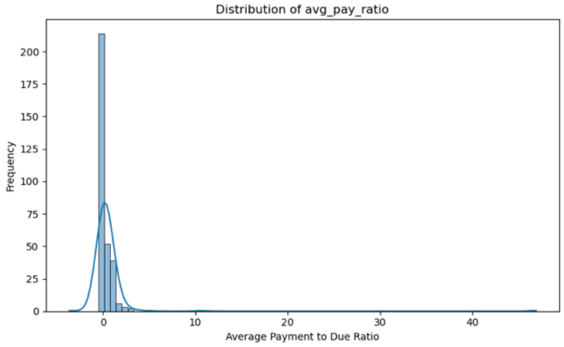
The analysis of the target variable “default” shows the distribution of customers who defaulted (19.4%) versus those who did not default on their credit card obligations (80.6%).

The left chart shows that customers with minimal delays (statuses -2, -1, 0) have low default rates between 5% and 12%, while default risk rises sharply with more severe delays, 41.7% at status 1 and 51.9% at status 2. This indicates that worsening payment behaviour is a clear early warning sign of default. The right chart supports this, showing that customers with continuous delinquency have a 47.4% default rate compared to just 17.6% for those without, highlighting that both the severity and consistency of late payments significantly increase credit risk.

The distribution of avg_pay_ratio is heavily right-skewed, with most customers paying little or only part of their dues, as shown by a sharp peak near zero. There are a few outliers pay well above the due amount, but they are not typical. Therefore, applying a log transformation may help to normalize the “avg_pay_ratio” variable.







Features such as “credit_limit”, “due_amt1”, “due_amt2”, “due_amt3”, “payment_amt1”, “payment_amt2” and “payment_amt3” (Appendix A) display strong right-skewness, suggesting most customers have relatively low balances or payments, while a small number have very large amounts, creating outliers. Therefore, applying a log transformation to these skewed features is recommended to normalize its distribution. Several features show strong correlations with each other, such as due_amt2, due_amt3, payment_status_1, payment_status_2, payment_status_3, payment_amt2, and payment_amt3 (*Appendix B*). Removing these highly correlated variables helps reduce multicollinearity and minimizes the risk of overfitting.

Model Evaluation

Six regression models were developed to predict the target variable. The dataset was split into 80% for training and 20% for testing to evaluate model performance, with 5-fold cross-validation applied during training. Model accuracy was assessed using AUC metric, allowing for comparison between in-sample and out-of-sample performance.

The baseline model, which uses all available features without selection or regularization, achieved an accuracy of 85.48% and a training AUC of 0.7593. However, the test AUC dropped significantly to 0.6288, indicating that the model is overfitting to the training data. This gap between training and test performance suggests the model captures noise or irrelevant patterns in the training set, resulting in poor generalization to unseen customers.

The forward selection logistic regression model chose a reduced feature set that included variables like credit_limit, payment_amt1, and worst_pay_status, among others. While it retained the same test accuracy of 85.48%, the AUC improved slightly to 0.6712, suggesting better risk ranking. The smaller feature set likely helped reduce overfitting, but some useful predictors might have been excluded, limiting further gains in performance.

The backward selection model also maintained the same accuracy but showed the lowest test AUC of 0.6019 despite a comparable training AUC. This further highlights that removing certain predictive features can hurt model generalization, and that the backward elimination process might have discarded variables that added subtle yet valuable information for distinguishing defaults.

The Lasso logistic regression model, which introduces L1 regularization to shrink and eliminate less useful coefficients, achieved a slightly lower accuracy of 83.87% but significantly outperformed all previous models on test AUC with a score of 0.8615. This strong AUC suggests that the Lasso model effectively reduced overfitting by simplifying the model while retaining the most important predictors for ranking credit risk. The lower training AUC also confirms that it avoids memorizing the training data.

Similarly, the Ridge regression model, which applies L2 regularization to shrink all coefficients without setting them to zero, showed balanced generalization. It produced a test AUC of 0.7962, better than the unregularized and feature-selected models, though not as high as Lasso or ElasticNet. Its slightly higher training AUC of 0.7533 indicates it retains more complexity than Lasso while still reducing variance.

The ElasticNet logistic regression model, which combines both L1 and L2 penalties, achieved the same performance as Lasso, with a test AUC of 0.8615 and accuracy of 83.87%. The optimal hyperparameters indicated a strong preference for L1 regularization (l1_ratio = 1.0), meaning the model effectively behaved like Lasso. Its strong test AUC and regularized structure confirm it is robust against overfitting while maintaining strong predictive power.

Model	Accuracy on Test Set	AUC on Training Set	AUC on Test Set
Standard Logistic Regression	0.8548	0.7593	0.6288
Forward Stepwise Selection	0.8548	0.7461	0.6712
Backward Stepwise Selection	0.8548	0.7585	0.6019
Lasso Regularization	0.8387	0.7282	0.8615
Ridge Regularization	0.8387	0.7533	0.7962
Elastic Net Regularization	0.8387	0.7282	0.8615

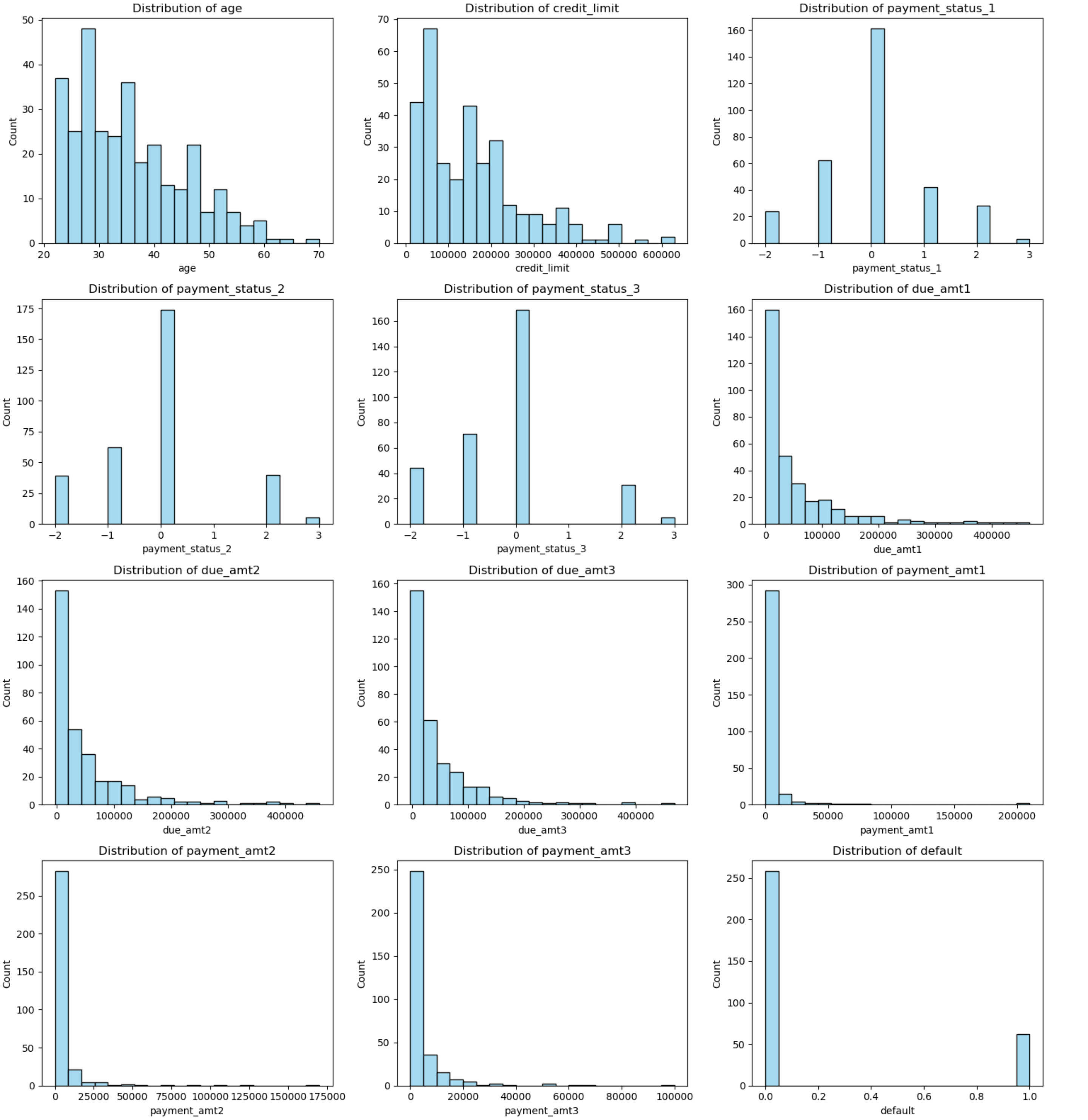
Table 1: Results Generated Across All Models

Conclusion

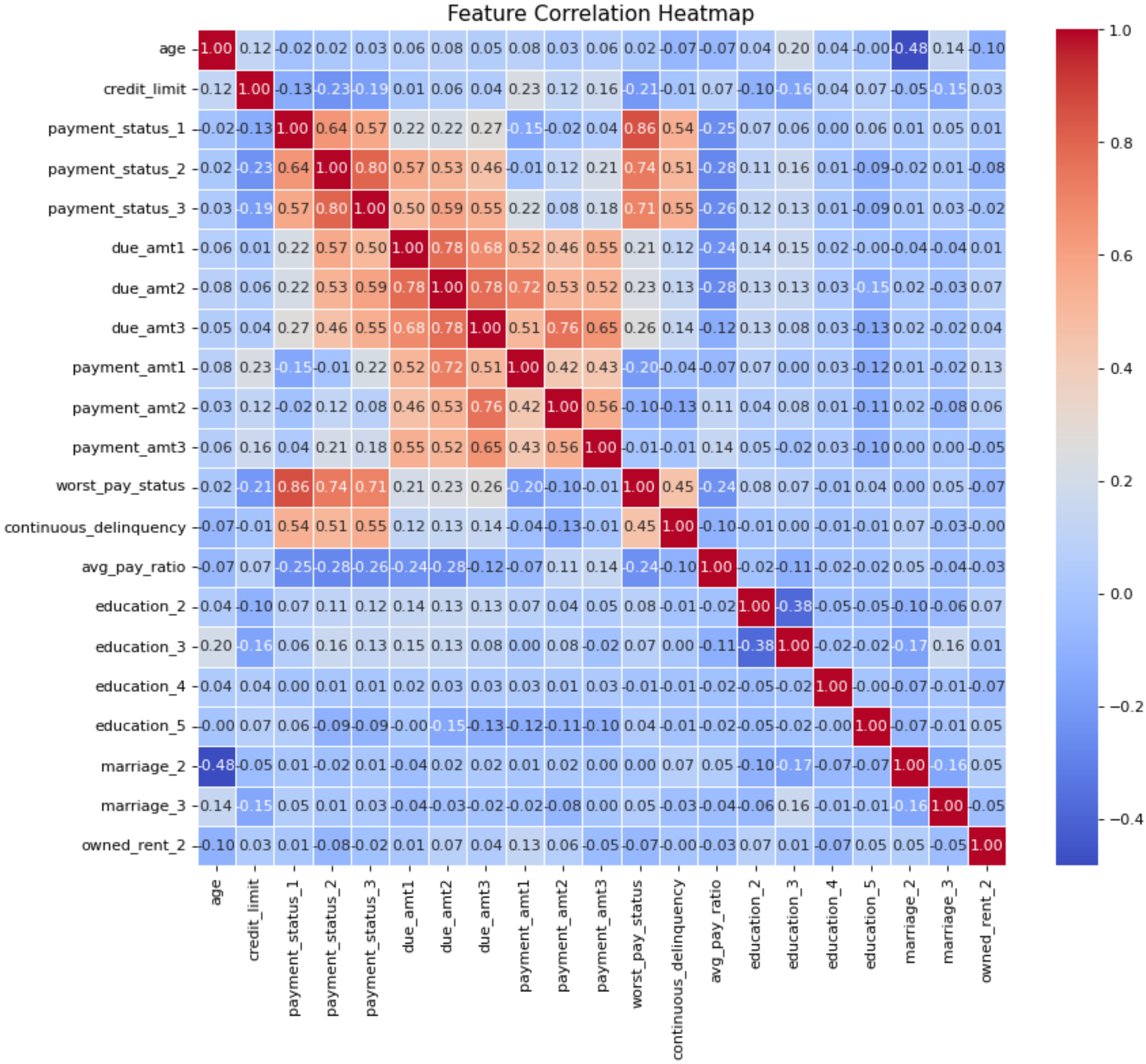
After evaluating all the models, the ElasticNet Logistic Regression model stands out as the most effective option for predicting credit risk in the bank's credit card portfolio. Although its test accuracy of 83.87% is slightly below that of the baseline and feature selection models, it achieves the highest test AUC of 0.8615. This indicates superior capability in distinguishing between defaulters and non-defaulters, which is particularly valuable for ranking customers by risk in a credit scoring application. The use of regularization in this model plays a crucial role in reducing overfitting and enhancing generalization to unseen data.

Appendices

Appendix A – Identify features to perform log transformation



Appendix B – Identify high correlated features to remove



	age	credit_limit	payment_status_1	payment_status_2	payment_status_3	due_amt1	due_amt2	due_amt3	payment_amt1	payment_amt2	payment_amt3	worst_pay_status	continuous_delinquency	avg_pay_ratio	education_2	education_3	education_4	education_5	marriage_2	marriage_3	owned_rent_2
age	1.00	0.12	-0.02	0.02	0.03	0.06	0.08	0.05	0.08	0.03	0.06	0.02	-0.07	-0.07	0.04	0.20	0.04	-0.00	-0.48	0.14	-0.10
credit_limit	0.12	1.00	-0.13	-0.23	-0.19	0.01	0.06	0.04	0.23	0.12	0.16	-0.21	-0.01	0.07	-0.10	-0.16	0.04	0.07	-0.05	-0.15	0.03
payment_status_1	-0.02	-0.13	1.00	0.64	0.57	0.22	0.22	0.27	-0.15	-0.02	0.04	0.86	0.54	-0.25	0.07	0.06	0.00	0.06	0.01	0.05	0.01
payment_status_2	0.02	-0.23	0.64	1.00	0.80	0.57	0.53	0.46	-0.01	0.12	0.21	0.74	0.51	-0.28	0.11	0.16	0.01	-0.09	-0.02	0.01	-0.08
payment_status_3	0.03	-0.19	0.57	0.80	1.00	0.50	0.59	0.55	0.22	0.08	0.18	0.71	0.55	-0.26	0.12	0.13	0.01	-0.09	0.01	0.03	-0.02
due_amt1	0.06	0.01	0.22	0.57	0.50	1.00	0.78	0.68	0.52	0.46	0.55	0.21	0.12	-0.24	0.14	0.15	0.02	-0.00	-0.04	-0.04	0.01
due_amt2	0.08	0.06	0.22	0.53	0.59	0.78	1.00	0.78	0.72	0.53	0.52	0.23	0.13	-0.28	0.13	0.13	0.03	-0.15	0.02	-0.03	0.07
due_amt3	0.05	0.04	0.27	0.46	0.55	0.68	0.78	1.00	0.51	0.76	0.65	0.26	0.14	-0.12	0.13	0.08	0.03	-0.13	0.02	-0.02	0.04
payment_amt1	0.08	0.23	-0.15	-0.01	0.22	0.52	0.72	0.51	1.00	0.42	0.43	-0.20	-0.04	-0.07	0.07	0.00	0.03	-0.12	0.01	-0.02	0.13
payment_amt2	0.03	0.12	-0.02	0.12	0.08	0.46	0.53	0.76	0.42	1.00	0.56	-0.10	-0.13	0.11	0.04	0.08	0.01	-0.11	0.02	-0.08	0.06
payment_amt3	0.06	0.16	0.04	0.21	0.18	0.55	0.52	0.65	0.43	0.56	1.00	-0.01	-0.01	0.14	0.05	-0.02	0.03	-0.10	0.00	0.00	-0.05
worst_pay_status	0.02	-0.21	0.86	0.74	0.71	0.21	0.23	0.26	-0.20	-0.10	-0.01	1.00	0.45	-0.24	0.08	0.07	-0.01	0.04	0.00	0.05	-0.07
continuous_delinquency	-0.07	-0.01	0.54	0.51	0.55	0.12	0.13	0.14	-0.04	-0.13	-0.01	0.45	1.00	-0.10	-0.01	0.00	-0.01	-0.01	0.07	-0.03	-0.00
avg_pay_ratio	-0.07	0.07	-0.25	-0.28	-0.26	-0.24	-0.28	-0.12	-0.07	0.11	0.14	-0.24	-0.10	1.00	-0.02	-0.11	-0.02	-0.02	0.05	-0.04	-0.03
education_2	0.04	-0.10	0.07	0.11	0.12	0.14	0.13	0.13	0.07	0.04	0.05	0.08	-0.01	-0.02	1.00	-0.38	-0.05	-0.05	-0.10	-0.06	0.07
education_3	0.20	-0.16	0.06	0.16	0.13	0.15	0.13	0.08	0.00	0.08	-0.02	0.07	0.00	-0.11	-0.38	1.00	-0.02	-0.02	-0.17	0.16	0.01
education_4	0.04	0.04	0.00	0.01	0.01	0.02	0.03	0.03	0.03	0.01	0.03	-0.01	-0.01	-0.02	-0.05	-0.02	1.00	-0.00	-0.07	-0.01	-0.07
education_5	-0.00	0.07	0.06	-0.09	-0.09	-0.00	-0.15	-0.13	-0.12	-0.11	-0.10	0.04	-0.01	-0.02	-0.05	-0.02	-0.00	1.00	-0.07	-0.01	0.05
marriage_2	-0.48	-0.05	0.01	-0.02	0.01	-0.04	0.02	0.02	0.01	0.02	0.00	0.00	0.07	0.05	-0.10	-0.17	-0.07	-0.07	1.00	-0.16	0.05
marriage_3	0.14	-0.15	0.05	0.01	0.03	-0.04	-0.03	-0.02	-0.02	-0.08	0.00	0.05	-0.03	-0.04	-0.06	0.16	-0.01	-0.01	-0.16	1.00	-0.05
owned_rent_2	-0.10	0.03	0.01	-0.08	-0.02	0.01	0.07	0.04	0.13	0.06	-0.05	-0.07	-0.00	-0.03	0.07	0.01	-0.07	0.05	0.05	-0.05	1.00