**Thesis Title** Finding Factors Affecting Marriage Rate and Marriage Prediction in China Using Panel Data Analysis and Machine Learning

**Author** Deyu Zhang

**Degree** Master of Science (Information Technology)

**Advisor** Asst. Prof. Worasak Rueangsirarak, Ph. D.

**Co-Advisor** Surapong Uttama, Ph. D.

**ABSTRACT**

After China’s accession to the WTO and 20 years of rapid development, the marriage rate has shown a downward trend. The main factors leading to the decline in marriage rate are the rapid growth of housing prices and the high price of betrothal gifts. Then, in this study, the adoption of big data analytic is proposed to highlight the significant factors effects to a decision making of new generation Chinese people.   
The first phase of research aims at fitting machine learning models with the marriage-related data, understanding which attributes affect the marriage rate and predicting the marriage rate. The data collection scope includes seven independent variables related to marriage rate such as GDP, house prices, birth rate, education level etc. over 31 regions in China during 2003-2022. Then the study applied three regression models - Pooled OLS, Random Effects, and Fixed Effects - in predicting China’s crude marriage rate. The Random Effects model outperformed both the Pooled OLS and Fixed Effects models, as evidenced by its highest R² value (0.2910). However, based on Hausman Test, p-value of 6.458e-16.the indicate Fixed Effects model was preferrable. All models suggested that the average year of education had the most positive effect to the marriage rate while the house price greatly negated the marriage rate. Results showed the Random Effects model, with an R² of 0.2910, as the best fit. Key predictors included GDP, house prices, and gross dependency ratio (negative effects), and sex ratio and education (positive effects). The Effects model excels in prediction, with the lowest MSE (1.6610), RMSE (1.2888) and Random Effects model excels in prediction, with the lowest MSE (1.6610), RMSE (1.2888) and MAE (1.0888).

The second phase of research study aims to analyze the impact of socio-economic factors on the crude marriage rate (CMR) panel data in China from 2003 to 2022 using Dual Machine Learning (DML) for Causal Inference and machine learning models. Four models—XGBoost, LightGBM, CatBoost, and GBDT—were employed for predictions, using 10-fold cross-validation for model evaluation. The results indicated that education and birth rate had the most significant positive impacts on CMR, while GDP showed positive but varying effects, and the female proportion had a notable negative impact. CatBoost performed best in MSE (0.942) and RMSE (0.958), while LightGBM excelled in MAE (0.777). Education, GDP, and birth rate are key factors influencing CMR. CatBoost and LightGBM proved to be effective prediction models, though improvements are needed for regions with significant variability.

After comparing different models, it can be concluded that the Random Effects model performed the best across all evaluation metrics (MSE, RMSE, MAE), demonstrating the advantage of traditional statistical models on this dataset. Although CatBoost performed relatively well among the machine learning models, its overall error was still higher than that of the Random Effects model, with XGBoost and GBDT showing larger errors. This indicates that, in this specific dataset, traditional statistical models outperform more complex machine learning models, highlighting the importance of optimizing model selection based on the characteristics of the data.

**Keywords:** Marriage Rate, Panel Data, Panel Regression, Hausman Test, Dual Machine Learning, Causal Inference, CatBoost, Marriage Rate Prediction

**Marriage**

China

Models and modelmaking

Forecasting-- **Marriage**