TIANHAO CAO

British Columbia | Email | Website | LinkedIn

OBJECTIVE

Machine Learning Engineer with a strong foundation in Business Management Economics (UCSC). Currently complete Master of Data Science (Computational Linguistics) at UBC, skilled in statistical analysis, machine learning, and causal inference. Eager to apply computational linguistics and data-driven insights to solve real-world challenges.

EDUCATIONAL BACKGROUND

University of British Columbia (M.S. in Data Science, Computational Linguistics) *September 2025 – June 2026*

- **Supervised Machine Learning, Unsupervise Machine Learning, Regression**

University of California, Santa Cruz (B.A. in Business Management Economic) *September 2019 – June 2022*

TECHNICAL SKILLS

Programming	Python, R, SQL, MongoDB
Data Analysis	Machine Learning, Neural Networks, LLMs, Feature Engineering, Data Visualization
Tools	PyTorch, Scikit-learn, XGBoost, Pandas, LaTeX, Git, Linux
Languages	English (Fluent), Mandarin (Native)

MAJOR PROJECTS

Credit Card Default Analysis Academic Project
University of British Columbia *GitHub Repository*

· Built an end-to-end machine learning pipeline to predict credit default probability using the UCI dataset (30,000 samples).
· Conducted extensive EDA to handle class imbalance and implemented feature scaling to improve model robustness.
· Trained and evaluated multiple classifiers (Logistic Regression, Random Forest, XGBoost), selecting the optimized XGBoost model which achieved the best balance of Precision and Recall (F1-score: 0.55).
· Utilized SHAP (SHapley Additive exPlanations) values to interpret black-box model decisions, identifying recent payment history as the most critical predictor of default.

Predictive Modeling of Agricultural Insurance Undergraduate Project
University of California, Santa Cruz *GitHub Repository*

· Engineered a robust dataset by cleaning and merging raw survey data from the Harvard Dataverse to analyze farmers' insurance purchase behavior.
· Developed and compared Lasso Regression (for featu re selection) and Random Forest models, achieving a peak predictive precision of 76%.
· Interpreted model outputs to identify key behavioral drivers, utilizing feature importance metrics to translate technical findings into actionable business insights.
· Conducted error analysis to assess prediction variance, discussing model limitations and potential biases in the survey data.

Causal Analysis of COVID-19 Policy Impacts Undergraduate Project
University of California, Santa Cruz *GitHub Repository*

· Aggregated and standardized cross-country economic panel data from the WHO and Our World in Data to construct a time-series dataset.
· Implemented a Difference-in-Differences (DiD) framework in R to quantify the causal impact of early pandemic lockdown policies on GDP across 5 major economies (US, China, Japan, UK, India).
· Mitigated omitted-variable bias by incorporating control variables, significantly strengthening the validity of the causal claims.
· Performed robustness checks by comparing the adjusted model against baselines to validate policy implications and discuss potential confounding factors.