Generative AI use statement

I, Yue Zhang, hereby confirm that I used chatgpt to polish my documentation. my team_contribution.md file. The output from the tool(s) was/were modified by further edited.

1. When using AI, I have ensured that the work produced is still my own and I understand that submitting unmodified output from a generative AI tool as my own is NOT acceptable. I understand that I am expected to build on the output, ensuring any submissions are my own ideas and knowledge.
2. I acknowledge awareness of any updates to the generative AI tools used, up to the date of this submission. This includes AI plug-ins or assistants included in existing programs, such as Grammarly. I take responsibility for any fabricated references or factual errors stemming from the use of these tools.
3. I have informed myself of the limitations and implications of using generative AI and related technologies, including the reinforcement of biases and propensity for fabrication.
4. I have used these tools ethically, including not uploading confidential, private, personal, copyrighted, or otherwise sensitive information.
5. To assist with maintaining academic integrity, I have appropriately acknowledged any use of generative AI in my work (list below as applicable).
6. I acknowledge that any undeclared use of generative AI will constitute academic dishonesty and will be dealt with according to relevant University policy.
7. I understand that I will be held accountable for any academic misconduct that arises in breach of any relevant University policy, as well as the consequences of such infringements.

Tool used:
chat gpt 5

Date accessed:
11 Oct, 2025

Prompt(s) entered:
I want you help detail based on below info

我建议一个人找linear 类型的model 像Ridge, 另一个找 tree style 类型的model 比较容易点因为本课成涉及较多像 randomforest, catboost, 虽然xuanhui 已经做了一些 tree 类型的 notebook

要不ming 找 tree 类型的model, zhengyang 找 linear 方向的model 吧 然后 xuanhui 主要负责立项，这个house project 就是他选的，还有找 外部的feature （cpi, gdp...）eda为什么 cpi gdp... can help etc. 同时也做model and parameter searching 我的话，就是做 infrastructure, data engineering, app deployment 的工作

becasue my project3 requriement says

Submit a separate document (or include this at the top of a notebook) that

> details what each team member contributed to the assignment. Not all
>
> contributors will be awarded the same mark. Each team member must
>
> submit their own version of how each team member contributed.
>
> so based on that could you write a formal english md file include all the
>
> information above
>
> thanks
>
> **Prompt output(s):**
>
> Project 3: Team Contributions
>
> # Team Members
>
> - Xuanhui
> - Ming
> - Zhengyang
>
> # What Each Person Did
>
> ## Xuanhui
>
> - Chose the house price prediction project
> - Found external data (CPI, GDP, OCR)
> - Did data analysis to show why these features help
> - Built and tested tree-based models
> - Tuned model parameters
>
> ## Ming
>
> - Built Random Forest and CatBoost models
> - Compared different tree-based models
> - Analyzed which features are most important
> - Optimized model settings
>
> ## Zhengyang
>
> - Built Ridge Regression and other linear models
> - Tested different regularization methods
> - Compared linear vs tree-based models
> - Selected best features for linear models
>
> # How We Split the Work

- **Tree Models**: Xuanhui and Ming worked on RandomForest, CatBoost, XGBoost
- **Linear Models**: Zhengyang worked on Ridge Regression
- **Infrastructure**: [Your Name] built the data system and app

This way everyone had their own focus and we covered different types of models.

Working Together

Everyone helped with:

- Team meetings
- Reviewing each other's code
- Final presentation
- Testing the system

Modification for assessment:

# Project 3: Team Contribu&ons ## Team Members

- **Xuanhui**

- **Ming**

- **Zhengyang**

- **Yue** (Infrastructure & Deployment Lead)

## Individual Contributions

### Xuanhui - Project Lead & Feature Engineering **Primary Responsibilities:**

- **Project Initiation**: Selected the New Zealand house price prediction project and defined the project scope

- **External Feature Research**: Identified and integrated external economic indicators (CPI, GDP, OCR) to enhance prediction accuracy

- **Exploratory Data Analysis (EDA)**: Conducted comprehensive analysis to justify the inclusion of macroeconomic features and their relationship with house price growth

- **Model Development**: Implemented and tested tree-based models with extensive hyperparameter tuning

> - **Model Optimization**: Performed systematic parameter search and optimization across multiple model architectures
>
> **Key Deliverables:**
>
> - Project proposal and scope definition
>
> - External data source integration (CPI, GDP, OCR datasets)
>
> - EDA notebooks demonstrating feature relationships
>
> - Tree-based model implementations with hyperparameter optimization
>
> ---
>
> ### Ming - Tree-Based Model Specialist **Primary Responsibilities:**
>
> - **Tree-Based Model Implementation**: Developed and optimized treebased models including Random Forest and CatBoost
>
> - **Model Comparison**: Conducted comparative analysis between different tree-based approaches
>
> - **Feature Importance Analysis**: Analyzed feature contributions specific to tree-based algorithms
>
> - **Performance Tuning**: Optimized hyperparameters for ensemble tree methods
>
> **Key Deliverables:**
>
> - Random Forest model implementation and evaluation
>
> - CatBoost model development (building on Xuanhui's initial work)
>
> - Tree-based model performance comparison documentation
>
> - Feature importance analysis for tree-based methods

> **Note**: While Xuanhui provided initial tree-based model notebooks, Ming expanded and specialized in this domain with additional models and deeper analysis.
>
> --- ### Zhengyang - Linear Model Specialist **Primary Responsibilities:**
>
> - **Linear Model Implementation**: Developed and optimized linear regression approaches including Ridge Regression
>
> - **Regularization Analysis**: Explored L1/L2 regularization techniques for linear models
>
> - **Model Comparison**: Compared linear approaches against tree-based methods
>
> - **Feature Selection**: Conducted feature selection specific to linear modeling assumptions
>
> **Key Deliverables:**
>
> - Ridge Regression implementation with cross-validation
>
> - Linear model performance evaluation and comparison
>
> - Analysis of feature linearity and multicollinearity
>
> - Documentation of linear vs. non-linear model trade-offs
>
> ---
>
> ### Yue - Infrastructure & Deployment Engineer **Primary Responsibilities:**
>
> - **Infrastructure Development**: Designed and implemented the complete data engineering pipeline
>
> - **ETL Pipeline**: Built automated Extract-Transform-Load system (`ingest_h.py`) for data ingestion from multiple sources
>
> - **Database Architecture**: Designed and implemented Supabase database schema (tables: `clean_house`, `feature_house`, `house_metrics`)

> - **Feature Engineering Pipeline**: Developed automated feature engineering module (`feature_engineering.py`) that generates 53 features from 8 base variables
>
> - **Model Training Framework**: Created unified training framework (`trainer.py`) with base class architecture for consistent model development
>
> - **Application Deployment**: Developed and deployed interactive Streamlit dashboard (`test.py`) for data exploration and model experimentation
>
> - **Cloud Deployment**: Deployed production application to Streamlit Cloud at [https://app-test-qxq5b9dukmh7yw6xuyufxc.streamlit.app/] (https://app-test-qxq5b9dukmh7yw6xuyufxc.streamlit.app/)
>
> - **DevOps**: Managed dependencies, environment configuration, and deployment pipeline
>
> **Key Deliverables:**
>
> - Complete ETL pipeline with multi-source data integration
>
> - Supabase database design and implementation
>
> - Automated feature engineering system with temporal features, lags, and rolling statistics
>
> - Unified model training framework supporting ETS, XGBoost, and CatBoost
>
> - Production-ready Streamlit dashboard with EDA tools and model training interface
>
> - Cloud deployment and production environment setup
>
> - Technical documentation and README
>
> **Technical Implementation:** - Implemented temporal feature engineering (lags: 1-4 quarters, 4 years; rolling means: 1yr, 4yr, 10yr)
>
> - Built modular trainer architecture enabling easy addition of new models
>
> - Integrated Plotly visualizations for interactive data exploration

> - Configured Supabase storage and database for team collaboration
>
> - Developed missing data handling strategies (drop/impute)
>
> ---
>
> ## Model Distribution Strategy The team adopted a complementary modeling approach:
>
> - **Tree-Based Models** (Ming, Xuanhui): RandomForest, CatBoost, XGBoost
>
> - leveraging course content and natural fit for tabular time-series data
>
> - **Linear Models** (Zhengyang): Ridge Regression and variants exploring regularization and interpretability
>
> This division ensured comprehensive model coverage while avoiding duplication of effort.
>
> ---
>
> ## Collaboration & Integration All team members contributed to:
>
> - Regular progress meetings and knowledge sharing
>
> - Code reviews and feedback
>
> - Documentation and final presentation preparation
>
> - Integrated system testing and validation
>
> The infrastructure developed by Yue served as the foundation enabling all team members to focus on their specialized modeling tasks with consistent data access and evaluation frameworks.