Thien Nguyen Hoang
Data Analyst

Khanh Hoa province • Viet Nam • nhthien30kh@gmail.com • 0868301039

Education

HCMC University of Transport HCMC, VN
Bachelor of Engineering in Electronics and Telecommunications. 09/2023

FPT Academy HCMC, VN
Automation Testing 2022

IIG HCMC, VN
Toeic 555 2023

Coursera HCMC, VN
Google Data Analytics, Google Advanced Data Analytics 2025

Projects

Chemical Exposure (DA)
- Built ETL pipeline for 25-year multi-source dataset (reducing 70%).
- Detected cross-network missing-pattern correlations between 2 monitoring systems.
- Revealing temporal anomaly clusters.

Cylistic Bike-share (DA)
- Processed and cleaned >1M trips, 92% improvement in data quality consistency.
- Identified casual riders take 66% longer trips.
- Revealing 50k-use high-traffic hubs near Lake Michigan and numerous low-activity stations.

Bella Beat Health (DA)
- Identified that ~60% of users fall into low–moderate activity.
- Detected abnormal HR patterns in ~8% of users.
- Estimated that targeted personalization could increase daily activity consistency by 12–18%.

Japan Restaurant (DA)
- Identified top 20% products contributing 80% revenue.
- Detected low-performing items (bottom 10%). Found peak hours.
- Found that promo users have +90% higher AOV.

Employee Churn Prediction (DS)
- Built an end-to-end automated ML pipeline (automated 85%).
- Identified the key churn drivers, high-risk employee groups.
- Improved model accuracy from 0.8 to 0.98 using XGBoost and deployed the final model with FastAPI.

Medical Cost Personal Prediction (DS)
- Built an end-to-end automated ML pipeline (automated 90%).
- Identified smoking as the dominant cost driver (+359% costs).
- Confirmed a positive linear relationship with age.

Skills

Databases: MySQL, BigQuery, Google Sheets, Excel
Visualization Tools: Tableau, Looker Studio, R, Plotly, Dash
Statistical Methods: Hypothesis Testing (t-test, chi-square), A/B testing, ANOVA
Languages: R, Python (Numpy, Pandas, Scipy, Seaborn, Matplotlib, Statsmodels, Scikit-learn)
Machine Learning Models: Regression (Linear, Logistic), Naive Bayes, Decision Trees, Random Forest, XGBoost.

Portfolio

https://github.com/nhthien306AuT/Portfolio_DA