Lulude Sun

luludesun@berkeley.edu | San Francisco Bay Area | U.S Citizen

EDUCATION

University of California, Berkeley   Dec. 2022

M.A. in Statistics
- **Courses:** Computational Data Science, Advanced Probability, Bayesian Statistics, Experimental Design, Linear Models

University of California, San Diego   Jun. 2021

B.S. in Applied Mathematics Minor in Machine Learning & Economics
- **Courses:** Machine Learning, Data Analytics, Computational Statistics, Business Analytics, Bayesian Inference

TECHNICAL SKILLS

- **Programming**: Python (Pandas, Scikit-Learn, Pytorch), SQL, R
- **Data Tools**: Tableau, Excel, Databricks, Pyspark, PowerBI, Git, LaTeX, SAS, GCP, AWS, Snowflakes
- **Data Science Methods**: A/B Testing, Data Wrangling, Database Management, Data Visualization, Time Series Forecasting, Machine Learning, Deep Learning

PROFESSIONAL EXPERIENCE

Data Scientist Intern   May – Aug. 2022

Reyes Coca-Cola Bottling   *Irvine, CA*

- Implemented **ETL solutions** with over 100GB data, created automated ML and data pipelines for inventory monitoring, optimized inventory levels for 2000+ machine parts, achieving over 100% improvement in inventory turnover rate.

- Conducted statistical analysis, built predictive models **ARIMA & XGBoost** in **Python** to forecast demand, reduced annual dollar waste in unnecessary inventory replenishment by 10% and enabling dynamic inventory threshold setting.

- Maintained data infrastructure and established model management system for continuous monitoring, incorporating real-time data feeds and automated alerts, and resolved data duplication problems using **Alteryx** and **Snowflake**.

- Constructed and tested data pipeline and fully automated & interactive **PowerBI** dashboards for internal stakeholders, streamlined inventory insights and provided grid-search-supported thresholds to warehouse operators.

Data Analyst Intern   Jul – Oct. 2020

H/A Studio   *San Francisco, CA*

- Designed and conducted **A/B testing** on user interface design, resulting in a 2% increase in conversion rate.

- Performed **SQL** queries to build data collection and statistical analysis using **ANOVA** in **Python** to evaluate the effectiveness of three different CTA button variations within the optimized UI framework.

- Implemented **Difference-in-Differences (DiD)** analysis using Python's **Statsmodels** to measure net effect of UI and CTA changes, revealing a net conversion increase of 1.1% attributable to the implemented changes.

PROJECTS

Advertisement CTR Prediction
- Analyzed user and ads data with Spark and predicted CTR using deep learning models in Pytorch.
- Built, trained, and evaluated variations of Wide and Deep neural network models; best of which reaches 74% AUC, 60% accuracy, 79% recall, and 26% precision.

Airbnb Demand and Revenue Prediction
- Analyzed Airbnb listings' demand and revenue in Los Angeles areas on Databricks with PySpark MLs.
- Surpassed baseline linear models by 30% RMSE with 10-fold cross-validation hypertuned XGBoost and Random Forest that predict occupancy rate and return-on-investment.
- Inferenced and developed insights from the model using SHAP to inform Airbnb hosts on optimal pricing strategy to achieve maximum return-on-investment.

Los Angeles Crime Rate Analysis
- Deployed SparkSQL and PySpark to perform data preprocessing, Exploratory Data Analysis, and spatial analysis from distributed datasets on LA crime rates to provide insights for law enforcements.
- Trained time series model with different time steps to model regional recurrent crime patterns.
- Navigated travelers with secure ways to destinations using machine learning predicted crime trends based on time and location to make travel and living recommendations.