

## Leo (Miao) You

linkedin.com/in/leoyouumd/

leoyou.netlify.app

github.com/leo-you



## TECHNICAL SKILLS

**Programming Language:** R, Python, SAS, VBA, PowerShell, HTML, Spark, SQL/T-SQL, NoSQL, MS SQL Server, MongoDB, Databricks, Azure

**Data Science, Machine Learning & Visualization:** EDA, ETL, SSIS, General Linear (Linear/Logistics/Polynomial/Regression with Regularization), SVM, Bayes, Tree-Based (Decision Tree with pruning, Random Forest, Boosted, etc.), K-NN, Clustering (K-MEANS, DBSCAN, Hierarchical), Hypothesis Testing, Google Analytics, Tableau, Power BI, SSRS, Advanced Excel (Formulas, Pivot, VBA, Power Query)

**Project Management & Version Control:** Agile, Scrum, TFS, Git, Confluence, Airflow, Docker

## WORK EXPERIENCE

### Circle K, Tempe, AZ, US - Sr. Data Scientist

2021 - Present

- Responsibilities include developing advanced analytics solutions and predictive models implemented in various areas of data-driven merchandising, designing and scaling up ETL pipelines to extract data from multiple sources within production environment, leveraged advanced machine learning techniques to build site selection tool that supports business units in identifying new sites etc.
- Led a team of 4 data scientists in developing non-linear optimization model in R using NLOPT framework, implemented localized pricing initiatives on 9 business units with over 3000+ stores and 20k+ items from data preparation, elasticity modeling, price optimization to recommend optimal price to key stakeholders, generating 1.2M+ annual margin uplift per business unit
- Designed automated ETL pipelines using PySpark and Databricks delta tables to extract data from multiple sources and transaction logs, modularized and scaled up legacy code in Python by encapsulating codes into classes and functions, reduced average runtime by 80%
- Built site selection tools using advanced ML models (Regularization, Tree-Based, XGBoost, etc.) based on 30+ features, automated model selection from 15+ models using Databricks AutoML, increasing model performance by 15% while reducing labor hour by 70%

### Center for Vein Restoration, Greenbelt, MD, US - Data Analyst

2016 - 2021

- Responsibilities include writing complex SQL queries for data extraction within Big Data environment, building fully automated ETL pipelines to ensure data quality, applying advanced ML techniques to increase business volume & revenue and drive decision making on opening new clinic locations, developing centralized reporting capabilities to satisfy analytical needs from all business levels, etc.
- Built fully automated ETL data pipelines with Python, REST API, and complex SQL queries joining and aggregating data from multiple sources within real time big data environment, applied advanced data cleaning techniques to handle missing values, outliers, distorted data distributions and ingest data under heavy workloads (10+MM) using Airflow scheduler, resulting in data accuracy improvement of 30%+ and manual labor hour cut down by 95+% (repurpose of 2+ FTEs)
- Leveraged advanced classification models (Logistic Regression, SVM, Tree-Based, etc.) to identify four most significant predictors that drive appointment booking rate. Fine-tuned model hyperparameters to increase accuracy by 20+% using Grid Search and Random Search. Model is being utilized in daily production and helped boost new patient booking rate by over 12%
- Conducted clinic segmentation analysis using Clustering (K-Means & Hierarchical) based on historical patient volume and revenue data, developed Decision Tree models with pruning on top of clustering results to predict clinic performance in undeveloped areas. Model results are being utilized by senior management as benchmarks to support decision making on opening new clinic sites
- Developed centralized interactive dashboarding capability using various reporting/BI tools such as SSRS and Power BI to track outbound recovery that are distributed as weekly KPI metrics for both leadership and frontline teams, with 150+ daily users and 100K+ total hits

### Envision Experience, Vienna, VA, US - Marketing Analyst

2015 - 2016

- Responsibilities include developing enterprise grade ETL data pipelines scraping and merging marketing campaign and website data from multiple sources, performing customer profiling analysis based on historical customer behavior patterns to drive social media advertising effectiveness and efficiency, implementing A/B testing using advanced statistical analysis to improve website UI design, etc.

## EDUCATION

### Harrisburg University of Science and Technology, Harrisburg, PA

2018

Master of Science in Analytics

### University of Maryland, College Park, MD

2015

Master of Science in Marketing Analytics

### Shanghai Normal University, Shanghai, China

2014

Bachelor of Engineering Major in Machinery Design & Manufacturing and Automation