2015

Strategic Planning Intern

Shin Yeong Project Management / Intern / Strategic Planning

- Researched competitors, customers, and legal issues for subsidiary establishment project.
- Created brochures, websites, and business cards for marketing and sales.
- Collected and organized occupational accident data into company database.
- Facilitated communication across planning, admin, marketing, design, and PM teams.
- Prepared real estate market outlook reports using population, GDP, and CPI indicators.
- Created 2 proposal reports and 5 market analysis documents.

Tech:

- Market Research
- Communication
- Cross-functional Collaboration
- MS Office

2017–2019

Military Police Sergeant

Republic of Korea Air Force / Sergeant / Military Police

- Led squad for base defense and search missions.
- Conducted security patrols and trained new recruits.
- Managed a team of over 15 personnel.
- Trained 15+ subordinates and ensured 0 security incidents.

Tech

- Leadership
- Security Operations

2019

Small and Medium-sized Enterprise(SME) Banking Intern

Korea Development Bank / Intern / SME Banking

- Conducted market research and analysis for SME financial evaluations.
- Assisted communication with overseas clients in English and Korean.
- Participated in due diligence and post-review of loan usage.
- Produced 2 market reports and assisted in 2 loan review projects.

Tech

- Excel, Market Research
- Communication

2021–2022

Project Manager

EXCELLO / Assistant Manager / Strategic Planning Department

- Joined EXCELLO as the 7th member and resigned after the company grew to a 29-member team.
- Partnered with world-renowned steel manufacturers to digitalize production equipment and enhance data-driven manufacturing processes. Defined optimization goals that resulted in a 10% cost reduction and a 25% improvement in energy efficiency, while preventing workplace accidents through data-driven safety management.
- Observed inefficiencies in traditional molten iron temperature measurement and handwritten data collection processes that had been used for decades, identifying the need for structured, digital data to support reliable analysis.
- Proposed a solution to improve the accuracy of temperature and refractory data collection, enabling data-driven decision-making in the steel production process.
- Self-taught hardware and software fundamentals to bridge communication gaps with engineers and business teams, enhancing collaboration efficiency across mechanical, electrical, software, and management departments.
- Gained deep insight into the transformative power of data in manufacturing and recognized the importance of data science as a foundation for future industrial innovation.
- Managed data-driven manufacturing production line digitalization “SMART” project for Ladle with Hyundai Steel.
- Managed data-driven manufacturing production line digitalization “SMART” projects for Runner, Ladle, and Tuyere Stock with POSCO.
- Managed data-driven manufacturing production line digitalization “SMART” projects for AOD and Continuous Casting Mold with SeAH Specialty Steel.
- Led AI-driven digital transformation national Research and Development (R&D) projects worth $25 million and served as project manager, overseeing planning, budgeting, proposal development, and project management from initiation in collaboration with stakeholders from government, corporations, and institutions.
- Provided data-driven business analysis and executive support to facilitate informed decisions in business operations and finance.
- Successfully managed over six large-scale technology projects and prepared more than fifteen comprehensive business proposals for corporate and government clients.

- Promoted from entry-level staff to Assistant Manager within seven months, in recognition of outstanding performance and successful project achievements.

Tech

- Tech Project Management
- R&D
- Energy Analytics
- Sales
- Marketing
- Public Relations
- Budgeting
- Finance
- Data Science
- AI & Industrial Optimization
- Manufacturing Digitalization
- Smart Factory
- Industrial IoT (IIoT)
- Predictive Maintenance
- Process Automation
- Sensor Data Acquisition
- Edge Computing
- Data Pipeline Architecture
- Time Series Analysis
- Energy Consumption Forecasting
- Process Optimization Modeling
- Data-driven Decision Making
- Cross-functional Team Leadership
- Technical Communication
- Stakeholder Management
- Systems Integration
- Digital Transformation Strategy
- Strategic Planning
- Sustainable Manufacturing
- Carbon Reduction Strategy
- Industrial Energy Analytics
- Energy Efficiency Optimization

Development of data-based energy efficiency optimization technology applicable to 70-ton Electric-Arc-Furnace in Steel making process

- Planned and proposed data-driven energy optimization solutions for Electric Arc Furnace (EAF) systems as part of a $12.5 million national R&D project in collaboration with 12 organizations, including government agencies, corporations, and research institutions, improving operational efficiency, reducing energy costs, and enhancing operator safety.

- Expected Outcomes:
  - ✓ Achieve cost reduction through energy efficiency optimization using AI-based predictive models for Electric Arc Furnace (EAF) operations.
  - ✓ Develop optimization models utilizing artificial intelligence to determine the optimal amount and timing of energy input.
  - ✓ Implement predictive operation models on-site to provide real-time operational guidance for furnace operators.
  - ✓ Establish an intelligent EAF operation system to derive optimal process conditions for improved productivity and energy savings.
  - ✓ Develop an integrated AI model that analyzes and manages the entire EAF process, moving beyond single-process monitoring.
  - ✓ Advance sensing technologies to acquire new types of process data for accurate AI-based optimization.
  - ✓ Enhance operator convenience and safety through AI-driven operational guidance, reducing reliance on manual experience.
  - ✓ Expand developed EAF predictive and optimization technologies across the steel industry to support digital transformation.
  - ✓ Contribute to the steel industry's transition toward carbon neutrality by improving EAF process efficiency and automation.
- Coordinated technical collaboration and budget allocation among participating institutions, managing a five-year national R&D project and facilitating the sharing of annual research outcomes across government, corporate, and academic partners.
- Consulted with experts on patent strategies to secure technological leadership and prevent information leakage, ensuring proactive protection and early positioning of core innovations.

Tech

- Artificial Intelligence (AI)
- Machine Learning (ML)
- Predictive Modeling
- Process Optimization
- Data-Driven Decision Making
- Time Series Analysis
- Energy Consumption Forecasting
- Anomaly Detection
- Reinforcement Learning for Process Control
- Manufacturing AI
- Sensor Data
- Energy Analytics

- Electric Arc Furnace (EAF) Optimization
- Smart Manufacturing
- Industrial IoT (IIoT)
- Process Automation
- Real-Time Operational Guidance
- Industrial Sensor Integration
- Edge Data Processing
- Digital Twin
- Energy Efficiency Optimization
- Energy Analytics
- Sustainable Manufacturing
- Carbon Neutral Transition
- Energy Cost Reduction
- Process Electrification
- Thermal Process Optimization
- National R&D Project Management
- Multi-Institutional Collaboration
- Budget & Resource Allocation
- Consortium Coordination
- Technology Commercialization Strategy
- Patent Strategy & Intellectual Property Protection
- Operator Safety Enhancement
- Real-Time Decision Support Systems

Development of combustion burner design/manufacture for 3MW and demonstration

- Served as Project Manager and lead organization representative for a national R&D project involving eight partner organizations, including corporations, research institutions, and government agencies.
- Led cross-institutional collaboration to develop a data-driven 3MW industrial combustion burner system aimed at enhancing thermal efficiency and reducing carbon emissions.
- Expected Outcomes:
  - ✓ Achieve significant reduction of greenhouse gas (GHG) emissions in reheating and forging furnaces using LNG or LPG by improving combustion efficiency and reducing fuel consumption.
  - ✓ Overcome the limitations of conventional combustion technologies that rely solely on waste-heat recovery by developing a high-efficiency oxy-fuel combustion system that minimizes CO_2 emissions.
  - ✓ Demonstrate the developed oxy-fuel combustion technology in large-scale steel manufacturing furnaces, enabling broader adoption across the steel industry.

  - ✓ Maximize thermal efficiency and energy utilization by replacing conventional air-fuel combustion with pure oxygen combustion.
  - ✓ Expand the application of oxy-fuel combustion technology to other industrial sectors such as heat treatment, ceramics, glass, and boiler manufacturing.
  - ✓ Contribute to national carbon neutrality goals by establishing the foundation for future hydrogen-based combustion systems, enabling Net-Zero industrial operations.
  - ✓ Enhance industrial competitiveness by reducing energy costs and enabling sustainable, energy-efficient manufacturing processes.
  - ✓ Provide a cost-effective CO_2 reduction solution for large-scale furnace operations, minimizing investment burden while maximizing environmental impact.
  - ✓ Establish oxy-fuel combustion as an alternative next-generation furnace technology for the steel industry.
  - ✓ Promote new investment, job creation, and industrial growth through commercialization and technology diffusion of advanced furnace systems.
- Coordinated technical collaboration and budget allocation among participating institutions, managing a five-year national R&D project and facilitating the sharing of annual research outcomes across government, corporate, and academic partners.
- Consulted with experts on patent strategies to secure technological leadership and prevent information leakage, ensuring proactive protection and early positioning of core innovations.

Tech

- Oxy-Fuel Combustion Technology
- Combustion System Design & Optimization
- Thermal Efficiency Enhancement
- Heat Transfer Analysis
- Fuel-Air Ratio Control
- Flame Stability Control
- Emission Reduction (CO_2, NOx)
- Carbon Emission Reduction
- High-Temperature Process Engineering
- Burner Prototype Development
- Predictive Modeling for Thermal Systems
- Machine Learning for Process Control
- Sensor Data Acquisition & Integration
- Energy Efficiency Optimization
- Low-Carbon Industrial Technology
- Hydrogen-Ready Combustion Systems
- Sustainable Manufacturing

- Carbon Neutrality Transition
- National R&D Project Management
- Multi-Institutional Collaboration (8 Organizations)
- Budget & Resource Allocation
- Consortium Coordination
- Patent Strategy & Intellectual Property Protection
- Technology Commercialization & Transfer
- Tech Project Management

2023–2025

Data Analyst

University of Southern California – Facilities Planning & Management (FPM)

- Analyzed university-wide utility bills and daily time-series water consumption data, focusing on irrigation efficiency and sustainability performance across campus facilities.
- Collaborated with the USC irrigation team to identify and categorize irrigation meters from all water usage meters, using Google Earth to map their respective meter locations and collect spatial area data for analysis.
- Developed an interactive irrigation usage dashboard, initially built with React.js and later migrated to Power BI, enabling daily monitoring and long-term trend analysis of water usage across all meter locations.
- Designed dashboard features to compare water consumption among similarly sized meter locations (4–5 per group), visualize monthly, quarterly, and yearly usage patterns, and highlight locations with 5%+ usage increases or highest consumption per square foot.
- Incorporated additional contextual data such as the presence of native Los Angeles plants, which influence irrigation demand, and delivered monthly insight reports to support sustainability-driven decision-making.
- Built an automated Python module to merge and update daily data from two separate water utility systems (without APIs) into a unified Excel-based database, ensuring standardized formatting and accessibility for non-technical staff.
- Enhanced the database with data quality diagnostics, including sheets that flagged missing data periods and zero-usage anomalies for each meter location.
- Enabled USC facilities managers and decision-makers to track irrigation performance, detect abnormal consumption patterns, and optimize water resource allocation, contributing to campus-wide sustainability and cost reduction goals.

Tech

- Python

- Pandas
- SQL
- React.js
- Power BI
- DAX
- MS Excel
- Excel Functions
- Google Earth
- Data Cleaning & Integration
- Time-Series Analysis (Daily)
- Anomaly Detection
- Data Quality Diagnostics
- Descriptive & Comparative Analytics
- Automated Data Processing (Python Module)
- Data Pipeline Automation
- Multi-System Data Consolidation
- Excel-Based Database Design
- ETL (Extract, Transform, Load) Workflow Design
- Google Earth Mapping
- Spatial Data Collection & Analysis
- Location-Based Resource Monitoring
- Irrigation & Utility Management Analytics
- Sustainability & Water Efficiency Analysis
- Power BI Dashboard Development
- React.js Interactive Dashboard
- KPI Visualization & Trend Analysis
- User-Driven Comparative Analysis
- Executive Decision Support Dashboards

Projects as Master's Student at USC

Forecasting Electricity Demand and Renewable Transition in Global Economies

- Predicted future electricity consumption, renewable energy adoption, and electricity prices across six major economies — China, the United States, India, Canada, and South Korea.
- Conducted multi-factor forecasting incorporating population growth, household expansion, and per-household electricity usage to estimate future total electricity demand.
- Modeled renewable energy generation ratios using historical data and estimated the technological development speed of renewables to project long-term transition rates.

- Predicted future electricity price per unit for both renewable and traditional energy sources, accounting for cost transition rates and country-specific energy efficiency gaps.
- Combined forecasts to calculate future national electricity expenditures, integrating renewable ratio (B), renewable price (C), and traditional price (D) into total consumption models.
- Collected and integrated open data from UN Data, U.S. Energy Information Administration (EIA), U.S. Department of Energy, National Bureau of Statistics of China, Ministry of Power (India), and Canadian Centre for Energy Information.
- Implemented data preprocessing and transformation pipelines to standardize datasets across heterogeneous international energy sources.
- Visualized forecasting results using bar, line, and ratio charts, illustrating cross-country comparisons in electricity usage, renewable adoption speed, and price evolution.
- Derived insights for energy policy and sustainability planning, highlighting potential electricity supply shortages and investment priorities in renewable infrastructure.

Tech

- Time-Series Forecasting
- Multi-Factor Regression Modeling
- Predictive Analytics
- Electricity Demand Modeling
- Renewable Energy Forecasting
- Energy Price Prediction
- Cross-Country Comparative Analysis
- Data Cleaning & Transformation
- Multi-Source Data Integration
- Data Pipeline Development
- Feature Engineering for Forecasting
- Handling Missing & Heterogeneous Data
- Linear & Polynomial Regression
- Model Validation & Error Analysis
- Global Energy Economics
- Renewable Energy Transition Analysis
- Sustainability Forecasting
- Carbon Neutrality Projections
- Energy Policy Insight Generation
- Data Visualization (Matplotlib, Seaborn)
- Trend & Ratio Analysis
- Cross-National Dashboard Design
- Insight Communication for Decision-Making

- Python
- Pandas
- NumPy
- Scikit-learn
- Statsmodels

Climate Impact on California Agriculture

- Examined how climate change affects California's agricultural economy, focusing on three major crops, almonds, grapes, and lettuce, which are strongly linked to California's GDP and agricultural exports.
- Collected climate data (temperature, precipitation, wind speed, snowfall) from the National Oceanic and Atmospheric Administration (NOAA) using API and agricultural price data from the Federal Reserve Bank (FRED) for the years 2008–2023.
- Selected four representative California regions, including Los Angeles, San Francisco, Yosemite, and Yreka, to capture diverse climate conditions across the state.
- Processed and merged datasets by aligning monthly agricultural price data with corresponding monthly averages of climate variables.
- Focused on the Producer Price Index (PPI) as the main dependent variable representing the economic value of each crop.
- Conducted data preprocessing including removing missing data, handling outliers, and cleaning inconsistencies across datasets.
- Applied data normalization and used average monthly temperature as a key independent variable to assess its correlation with agricultural prices.
- Introduced a 3-month lag to capture delayed climate effects on crop production and market prices.
- Performed linear regression analysis to evaluate the relationship between temperature and PPI for each crop.
- Found that the correlation between temperature and crop PPI increased when applying the 3-month lag adjustment, especially for lettuce.
- Developed a Long Short-Term Memory (LSTM) model to predict monthly PPI values using normalized time-series data.
- Compared the LSTM model performance with the linear regression model, using metrics such as Mean Squared Error (MSE) and R^2.
- Observed that the LSTM model provided more accurate forecasts and better captured seasonal fluctuations than the linear regression model.
- Determined that lettuce showed the highest model accuracy and strongest climate sensitivity, while almonds and grapes exhibited smaller variations.
- Visualized historical and forecasted PPI trends for all three crops using Matplotlib, comparing actual versus predicted values.

- Highlighted the impact of temperature as a major influencing factor on short-term agricultural price movements in California.
- Concluded that data-driven forecasting methods such as LSTM can be used to anticipate agricultural market trends under changing climate conditions.
- Suggested that ongoing climate monitoring and predictive modeling could help policymakers and producers adapt to future climate variability and maintain agricultural stability.

Tech

- Python
- Pandas
- NumPy
- Scikit-learn
- TensorFlow / Keras (LSTM)
- Statsmodels
- National Oceanic and Atmospheric Administration (NOAA) Data / API
- Federal Reserve Bank (FRED) Data
- Data Cleaning & Feature Engineering
- Time-Series Forecasting
- Linear Regression
- LSTM Neural Networks
- Climate Data Analysis
- Agricultural Economics
- Producer Price Index (PPI) Modeling
- Visualization & Trend Analysis

Web-Based Application for U.S. Presidential Election Sentiment Analysis Using Reddit Data

- Developed a web-based application that performs sentiment analysis on Reddit posts and comments to study public opinion related to the U.S. presidential election.
- Implemented using Flask (Python) for backend, React.js for frontend, and Google Firebase Realtime Database for data storage and management.
- Managed project scheduling, data collection, database management, sentiment analysis, and report writing collaboratively as part of a 3-member team.
- Initially planned to collect Twitter data via API, but due to policy and access limitations, shifted to Reddit as the primary data source.
- Used PRAW (Python Reddit API Wrapper) to scrape Reddit posts and comments from politically relevant subreddits, capturing titles, comment text, user IDs, timestamps, upvotes, downvotes, and URLs.
- Collected and stored over 7,000 comments from Reddit across multiple subreddits.

- Utilized Google Firebase Realtime Database (NoSQL) with partitioning, replication, and horizontal scaling strategies to distribute and balance data across four distinct databases.
- Implemented hash-based data partitioning, using ASCII-sum hashing to evenly distribute Reddit post data and improve load balancing and system performance.
- Structured all data in JSON format to enable real-time synchronization between Firebase and the web interface.
- Developed extract_data.py to automate the data scraping and insertion into Firebase in real time.
- Integrated Flask backend routes for CRUD operations (Create, Read, Update, Delete), enabling users to insert, edit, or delete posts through the web interface.
- Designed a search functionality for keyword, subreddit, and metadata-based queries to filter and view posts dynamically.
- Used NLTK (Natural Language Toolkit) in Python to perform sentiment analysis on each Reddit comment.
- Calculated a sentiment score between -1 and 1 for each comment, where values near 1 represent positive sentiment, near -1 represent negative sentiment, and 0 indicates neutrality.
- Processed keyword-based sentiment scores for four political entities — Trump, Biden, Republican, and Democrat — by aggregating positive and negative scores.
- Created score_parse.py with functions calculate_average and calculate_percentage_difference to compute average positive/negative scores and normalize support ratios between candidates and parties.
- Combined candidate and party sentiment ratios into a unified support score metric, representing total public sentiment share.
- Predicted the election outcome based on these computed support scores, determining a lead margin of 0.38% in favor of Trump at the time of analysis.
- Displayed real-time visual updates of Reddit sentiment trends on the web interface via React.js components.
- Ensured real-time updates between users and Firebase using automatic propagation and two-way synchronization.
- Implemented secure and scalable database integration that maintains high availability and fast read/write operations for all user interactions.
- Designed an intuitive UI that allows users to view data in a tabular format rather than raw JSON, improving usability and analysis accessibility.
- Encountered challenges with Firebase's monitoring and analytics tools being available only under paid subscriptions; resolved by developing custom Python scripts for database health checks.
- Adapted project direction to overcome limited API access and paid tool restrictions while maintaining real-time functionality and performance.
- Collected approximately 1,900 analyzed comments used to compute sentiment scores and predictions.

- Evaluated the application’s scalability and responsiveness, ensuring performance efficiency through distributed data architecture.
- Proposed future enhancements, including integrating Twitter data and automating model updates every 3 hours via scheduled jobs (cron).
- Concluded that social media sentiment analysis can effectively model and predict public opinion trends using open data, demonstrating a practical case of data science in political analytics.

Tech

- Python
- Flask
- React.js
- Firebase Realtime Database (NoSQL)
- PRAW (Python Reddit API Wrapper)
- NLTK (Sentiment Analysis)
- Data Partitioning · Replication · Horizontal Scaling
- JSON Data Structure
- CRUD API Development
- Real-Time Data Synchronization
- Hash-Based Data Distribution
- Web Application Development
- Sentiment Score Computation
- Data Visualization & Dashboard
- Election Sentiment Forecasting
- Cross-Platform Data Integration
- Automation & Scheduling (Python Script)

Sentiment Classification of Movie Reviews Using Deep Learning Models

- Built a sentiment classification model to distinguish between positive and negative movie reviews using the IMDB dataset.
- Implemented the project in Python (Jupyter Notebook) with Keras, TensorFlow, NumPy, Matplotlib, and NLTK libraries.
- Labeled reviews as positive or negative and loaded text data from local directories.
- Cleaned text data by removing punctuation, numbers, and special characters with regular expressions.
- Tokenized reviews using Keras Tokenizer and converted them into integer sequences based on word frequency.
- Calculated dataset statistics such as vocabulary size and average review length, and visualized review length distribution to determine sequence limits.
- Applied truncation and zero-padding to create uniform input sequences for deep learning models.

- Created a 32-dimensional embedding layer to transform words into dense numerical vectors.
- Built and trained a Multi-Layer Perceptron (MLP) model with multiple dense layers and dropout regularization.
- Implemented a 1D Convolutional Neural Network (Conv1D) with convolution and pooling layers to extract local word pattern features.
- Designed a Long Short-Term Memory (LSTM) network to capture sequential dependencies and contextual meaning in review texts.
- Applied dropout regularization, batch size, and epoch settings to improve model generalization and prevent overfitting.
- Used binary cross-entropy as the loss function and Adam optimizer for model training.
- Evaluated all models using accuracy, mean squared error, and loss metrics to assess performance.
- Visualized model training results through accuracy and loss plots using Matplotlib.
- Compared MLP, CNN, and LSTM architectures, confirming that LSTM achieved the best generalization by effectively capturing sequential text dependencies.
- Concluded that recurrent neural networks are the most suitable for sentiment classification tasks involving natural language data.

Tech

- Python
- Keras
- TensorFlow
- NumPy
- Matplotlib
- NLTK
- Text Cleaning & Preprocessing
- Tokenization
- Word Embedding
- Multi-Layer Perceptron (MLP)
- Convolutional Neural Network (Conv1D)
- Long Short-Term Memory (LSTM)
- Dropout Regularization
- Binary Cross-Entropy Loss
- Adam Optimizer
- Model Evaluation & Visualization
- Sentiment Classification
- Deep Learning for NLP

XGBoost-Based Recommendation System for Yelp Data Using Spark RDD

- Built a hybrid recommendation system combining item-based collaborative filtering (CF) and model-based regression using XGBoost, fully implemented with PySpark RDDs to meet distributed processing and efficiency requirements of the competition.
- Improved upon the Assignment 3 baseline by integrating feature-rich model-based learning with CF predictions to reduce RMSE and enhance generalization.
- Processed the official Yelp Dataset Challenge data (user.json, business.json, review_train.json, yelp_train.csv) using Spark RDD transformations for large-scale distributed data handling.
- Loaded and filtered the yelp_train.csv data to extract (user_id, business_id, rating) triples for training and validation.
- Parsed JSON data using SparkContext.textFile() and json.loads() for efficient parallel deserialization of millions of records.
- Engineered user-level features such as average star rating, review count, fan count, and usefulness score from user.json.
- Engineered business-level features such as average stars, review count, categories, and nested JSON attributes like parking availability from business.json.
- Extracted review-level features — average "useful," "funny," and "cool" scores per business — by grouping and averaging records from review_train.json.
- Encoded categorical data (e.g., number of categories) numerically and normalized continuous features.
- Implemented a feature-imputation routine to replace missing values with neutral defaults (e.g., mean = 0 or 3.5) to ensure model stability.
- Constructed training and validation matrices (X_train, Y_train) by joining user, business, and review dictionaries for each (user_id, business_id) pair.
- Used Numpy arrays for feature input to the machine-learning component while retaining Spark RDDs for preprocessing and distributed data loading.
- Trained the XGBoost regression model (XGBRegressor) using tuned hyperparameters including regularization (lambda, alpha), subsampling, maximum depth, and learning rate.
- Leveraged GPU-independent XGBoost configurations optimized for CPU and memory efficiency under Spark execution constraints.
- Used binary regression output to predict continuous star ratings between 1 and 5.
- Implemented item-based collaborative filtering (CF) using Pearson correlation similarity to model user preference similarity between co-rated businesses.
- Designed weighted-average rating prediction using top-N most similar items for each user-business pair, applying confidence-based smoothing for sparse cases.
- Added fallback logic for users or businesses with few co-ratings, defaulting to user or business mean ratings.
- Cached intermediate dictionaries (bus_user_dict, user_bus_dict, bus_user_r_dict, bus_avg_dict) to avoid redundant computation and reduce Spark shuffle time.

- Combined model-based and CF outputs into a hybrid prediction using an adjustable weighting factor (factor = 0.001) to minimize RMSE.
- Experimented with multiple hybrid weights (0.0001 – 0.2) and selected the value yielding the largest RMSE drop without additional runtime overhead.
- Validated model performance on yelp_val.csv, producing final predictions in CSV format with header user_id,business_id,prediction.
- Implemented comprehensive error distribution analysis, classifying prediction deviations into <1, 1–2, 2–3, 3–4, >4 bins to evaluate stability and bias.
- Calculated and reported RMSE on validation data (≈ 0.981) and total execution time (~6 minutes), demonstrating competitive leaderboard performance.
- Optimized Spark job efficiency via RDD caching, avoiding broadcast overloads, and minimizing JSON parsing repetition.
- Included detailed inline documentation in competition_w_comment.py summarizing methodology, parameter tuning, and runtime metrics per competition rules.
- Delivered a reproducible, scalable, and interpretable pipeline integrating distributed data processing (Spark) and machine-learning regression (XGBoost) for personalized rating prediction at industrial scale.

Tech

- Python
- PySpark (RDD)
- XGBoost (Model-Based Regression)
- Item-Based Collaborative Filtering (CF)
- Pearson Correlation Similarity
- Hybrid Recommendation System
- Distributed Data Processing
- Feature Engineering
- JSON Parsing & Aggregation
- Data Imputation
- RMSE Evaluation & Error Distribution
- Model Tuning & Regularization
- Weighted Hybrid Prediction
- Recommender System Optimization
- Scalable Machine Learning

Neighborhood Visualization and Analysis: A Guide to Los Angeles for Travelers and Residents

- Developed an interactive web-based visualization platform that integrates diverse urban datasets — including crime reports, hotel reviews, restaurant distributions, and bike-sharing infrastructure — to help both travelers and residents make informed decisions about neighborhood safety, amenities, and accessibility in Los Angeles, particularly in preparation for global events such as the Olympic Games and FIFA World Cup.

- Conducted extensive data acquisition and integration from reliable open sources, collecting LAPD crime reports (2020–2023) via Kaggle with details on incident type, time, and location; Booking.com hotel review data containing cleanliness, safety, and value ratings aggregated by ZIP code; Yelp restaurant data including geospatial coordinates, categories, and average ratings for density-based visualization; and Metro Bike Share trip data covering trip duration, start and end stations, and route paths for mobility analysis.
- Cleaned and standardized all datasets using Python (pandas, geopandas), converting them into unified geospatial formats (GeoJSON, CSV), aligning coordinate systems, resolving projection inconsistencies, and aggregating neighborhood-level statistics for accurate spatial visualization.
- Designed the prototype layout in Figma, outlining core sections such as an interactive 3D map panel, sidebar filters for dataset selection, and dynamic chart components for comparative analysis across neighborhoods.
- Implemented the front-end using Vue.js for reactive component-based rendering and modular structure, ensuring real-time interactivity and responsiveness.
- Integrated Deck.gl for 3D geospatial visualization, employing hexagonal bin layers to represent crime density, restaurant concentration, and hotel distribution, with adjustable parameters for color scale, radius, and opacity to enhance clarity and insight.
- Combined Mapbox as a base map provider for geographic context and D3.js for interactive analytical charts displaying temporal trends, categorical distributions, and neighborhood-level metrics.
- Created intuitive user interactions where clicking on a neighborhood reveals localized analytics such as crime frequency by type, hotel review breakdowns, and restaurant density, while hover actions display tooltips with contextual summaries.
- Enabled dynamic filtering and comparison of neighborhoods based on multiple dimensions — such as crime rates, amenity quality, and accessibility — allowing users to tailor insights to their personal safety or convenience preferences.
- Integrated bike-sharing route visualization to highlight safe and efficient cycling corridors, overlaying bike paths with crime data to assist in identifying low-risk commuting options.
- Applied data aggregation and clustering algorithms to simplify visualization of high-density areas, particularly in central and tourist-heavy zones like Downtown and Hollywood, improving usability and map readability.
- Leveraged generative AI tools such as ChatGPT and Copilot to troubleshoot Deck.gl–Vue.js rendering conflicts, optimize performance during multi-layer rendering, and enhance user interface alignment across components.

- Addressed technical challenges related to coordinate misalignment, dataset heterogeneity, and inconsistent temporal granularity, ensuring synchronized and reliable cross-layer analytics.
- Optimized performance by caching geospatial data, implementing client-side filtering, and minimizing re-rendering overhead to maintain interactive frame rates on large datasets.
- Designed use-case scenarios that demonstrated the platform's ability to support travel planning, such as identifying safe neighborhoods with high hotel satisfaction and dense restaurant options, and daily urban decision-making by residents through data-driven visualization.
- Highlighted eco-friendly mobility by visualizing bicycle infrastructure and promoting sustainable transportation routes through data overlays combining safety, accessibility, and convenience metrics.
- Delivered a cohesive, high-performance, and visually engaging web platform integrating geospatial data science and front-end engineering to provide actionable insights into the dynamics of Los Angeles neighborhoods.

Tech

- Python
- Pandas
- GeoPandas
- Vue.js
- Deck.gl
- D3.js
- Mapbox
- Figma
- JavaScript (Frontend Integration)
- Geospatial Data Visualization
- Data Cleaning & Standardization
- Interactive Dashboard Development
- Crime & Safety Analytics
- Restaurant Density Mapping
- Hotel Review Aggregation
- Bike Route Visualization
- Marker Clustering & Aggregation
- Urban Mobility Data Integration
- Generative AI for Debugging
- 3D Hexagon Layer Visualization
- Sustainable Transportation Analysis

Investigating Bias in AI for Sepsis Diagnosis: Analyzing Ethnic Disparities in ICU Decision-Making

- Conducted a machine learning–based fairness analysis using the MIMIC-III critical care database to investigate racial and ethnic disparities in AI-assisted sepsis diagnosis within intensive care unit (ICU) environments, focusing on bias in model decision-making processes.
- Defined sepsis-positive and sepsis-negative cohorts using ICD-9 diagnosis codes 99591, 99592, and 78552 to ensure clinical validity, and filtered the dataset to include only adult patients with complete demographic information such as age, gender, ethnicity, and insurance type.
- Performed data preprocessing using Python libraries including pandas, scikit-learn, and imbalanced-learn by applying one-hot encoding to categorical variables, min–max normalization to numerical attributes, and random oversampling to balance sepsis-positive and sepsis-negative samples.
- Split the dataset into training and testing subsets in an 80/20 stratified manner to maintain proportional class distribution and prevent sampling bias across all ethnic and demographic groups.
- Developed three predictive models, Logistic Regression, Random Forest, and Multi-Layer Perceptron (MLP), to compare differences in accuracy, interpretability, and fairness across models of varying complexity.
- Configured logistic regression with L2 regularization for baseline interpretability, tuned Random Forest with controlled depth and tree count for ensemble-based prediction stability, and implemented an MLP neural network with two ReLU-based hidden layers optimized using the Adam optimizer and binary cross-entropy loss.
- Evaluated predictive performance using metrics such as AUROC, precision, recall, F1-score, and confusion matrices, confirming Random Forest achieved the best balance between sensitivity and specificity.
- Performed statistical hypothesis testing including chi-square analysis to confirm significant relationships between ethnicity and sepsis diagnosis rates, establishing empirical evidence for bias presence before modeling.
- Applied SHAP (SHapley Additive Explanations) to interpret model predictions and quantify the contribution of each input feature to the final classification outcomes, providing transparency into how models weighted demographic versus clinical factors.
- Identified that non-clinical features such as ethnicity, insurance type, and age had disproportionately high SHAP values across all models, revealing that social and demographic variables strongly influenced AI-based sepsis predictions.
- Calculated fairness metrics including False Negative Rate (FNR) and False Positive Rate (FPR) for each ethnic group to quantify differences in diagnostic error patterns and detect underdiagnosis or overdiagnosis tendencies.

- Found that Black, Hispanic, and Middle Eastern patients exhibited higher FNR values indicating underdiagnosis risk, while Asian subgroups showed elevated FPR values indicating a higher likelihood of false positive predictions.
- Demonstrated that model complexity did not inherently improve fairness since Random Forest and MLP models showed higher accuracy but similar or greater disparity across demographic groups compared to logistic regression.
- Implemented bias mitigation through adversarial debiasing by training a secondary classifier to minimize the predictability of protected attributes from latent representations and through post-hoc threshold adjustment by setting group-specific decision thresholds to equalize error rates.
- Compared mitigation results using Equalized Odds metrics and found threshold adjustment reduced FNR disparity to 0.05 while adversarial debiasing reduced FPR disparity to 0.13, confirming each method's partial effectiveness in fairness improvement.
- Conducted SHAP interaction analysis to explore interdependencies among variables such as the combined effect of age and insurance type, revealing structural socioeconomic biases embedded in prediction mechanisms.
- Visualized fairness comparisons through FNR and FPR bar plots and SHAP-based heatmaps that depicted intergroup disparities in model reasoning and identified which variables most contributed to bias.
- Determined that threshold calibration produced the most interpretable and practically deployable mitigation strategy for clinical environments because it improved fairness without significantly compromising predictive power.
- Concluded that accuracy optimization alone is insufficient for equitable clinical AI systems and emphasized the importance of incorporating fairness constraints into the training process of diagnostic models to reduce systemic healthcare disparities.
- Documented the entire analysis pipeline in the accompanying Jupyter Notebook environment including data extraction, preprocessing, model training, SHAP interpretation, bias quantification, and visualization of mitigation effects to ensure transparency and reproducibility of results.

Tech

- Python
- Pandas
- NumPy
- Scikit-learn
- Imbalanced-learn
- Logistic Regression
- Random Forest
- Multi-Layer Perceptron (MLP)
- SHAP
- Explainable AI (XAI)

- Fairness in Machine Learning
- Bias Mitigation
- Adversarial Debiasing
- Threshold Adjustment
- Equalized Odds
- AUROC
- Precision
- Recall
- F1-score
- Statistical Testing
- Feature Importance
- MIMIC-III Database
- Sepsis Diagnosis
- Responsible AI
- Healthcare Analytics