✉ mindyng8855@gmail.com
🌐 mindyng.github.io
☎ 5105082455
in
www.linkedin.com/in/mindyng85
mindyng

I am passionate about combining descriptive analytics with results-oriented data problem-solving and bridging the knowledge gap across multiple disciplines and presenting insights/results to different audiences and teams.

Skills

PROJECT MANAGEMENT

Scoping out Business Problem
Defining Project Success
Metrics Development
Defining KPI's
Team-Player
Cross-Discipline Collaboration
Insights to Stakeholders

LANGUAGES

SQL
Python

DATA ENGINEERING (ELT)

Snowflake
Meltano
PostgreSQL DB
dbt
Dataform

DATA WRANGLING

Data Cleaning
Data Normalization
Data Integrity Checks / Assertions

STATISTICS

Descriptive Statistics
Probability Statistics
Inferential Analytics
Hypothesis Testing
A/B Testing

MODELS / MACHINE LEARNING

Linear Regression
Logistic Regression
Decision Trees
Random Forests
Naive Bayes Classification
K-Means Clustering
Natural Language Processing (NLP)

BUSINESS ANALYTICS

Cohort Analysis
Time Series Analysis
Churn Prediction

VISUALIZATION / BUSINESS INTELLIGENCE

Looker
Superset
Tableau

MINDY NG

DATA ANALYST

Projects

Deployed Web App for Business Stakeholder June 2021 to June 2021

Took a business stakeholder's question, translated it into a data question, performed ETL and created a web app with visualizations and analytic conclusion. Deployed web app with non-technical language using Streamlit's Sharing feature.

Modern Analytics Data Stack Built from Scratch Feb. 2021 to Feb. 2021

Built from the ground up and maintaining- data pipeline to perform ELT and BI visualization layer. Stack included open source: PostgreSQL DB, Meltano, dbt and Superset.

Music Streaming Service Churn Prediction Jan. 2021 to Jan. 2021

543,705 samples of user data used to investigate what leads to churn and to predict its occurrence.

Best model (Logistic Regression) had f1-score of 0.5 for minority class.

Model can be used to foresee which customers are likely to cancel their subscription so business can intervene to maintain high revenue stream.

Healthcare Workers' Burnout Classifier Jan. 2021 to Jan. 2021

Scraped 1879 tweets from nurses on the front lines in order to build a sentiment classifier to predict burnout.

Best model (LSTM) had f1-score of .51 for minority class before deployment using Streamlit.

Web application can be used for hospital directors to intervene on burnout to sustain healthcare workforce.

Time Series Forecasting on Uber Eats' Vendors Dec. 2018 to Dec. 2018

Utilized 7,911 samples of date-stamped data and predicted which vendors were worth continuing business with based on ROI.

Trended each vendors' data with Facebook's Prophet. Trends performed over a span of 15 months. Data further broken down into weekly and daily trends. Resulting model performance based on 30-day horizon producing 0.01 - 0.03 RMSE.

Postmates New Market Analysis with Geospatial Heatmaps Mar. 2019 to Mar. 2019

Analyzed 3-sided market to explore contributors to conversion and churn, used heatmaps to visualize supply and demand, determined health of market and addressed data integrity issues.

Sentiment Classification on Amazon Book Reviews Feb. 2017 to Apr. 2017

Gathered 243,269 Amazon book reviews through UCI's Machine Learning Repository in order to label customer reviews with three different sentiment scores to allow efficient product assessment.

Built three different classification models- MN Naive Bayes, Decision Tree and Random Forest.

Out of the three, Random Forest was the best predictor due to having best model performance results with 0.72 Test Set Accuracy. Reclassifying Amazon product reviews prevents shopping paralysis leading to quick purchase conversions.

Medicare Prescription Drugs Analysis July 2019 to July 2019

Analyzed 25,209,130 samples of Medicare Part D Prescription use to determine how geography correlates with provider density, provider specialties and drug costs.

Plotly and Seaborn used to visualize number of providers across states, to geocode provider specialties and to examine differing degrees of drug cost variance across the U.S.

Cohort Analysis on Drugs for Cancer Patients Jan. 2019 to Jan. 2019

Examined 1,096 samples of de-identified cancer patient treatment data to predict best drug regimen for cancer clinic's cohort.

Utilized paired t-test to determine if there was difference in efficacy between two different Breast Cancer drugs.

Employment

Mercari US Palo Alto, CA
Data Scientist Sept. 2022 to Oct. 2022, Sept. 2022 to Oct. 2022

- Helped build out search team's new metrics table in Dataform to power dashboards and assess experiment results.
- Analyzed price elasticity between highest GMV/most searched for category groups to understand buyers' demand in relation to item price changes.

Mercari US Palo Alto, CA
Business Intelligence Analyst Aug. 2021 to Aug. 2022

- Created Search RFM segmentation to personalize search.
- Investigated query chaining to understand search engine performance as well as searchers' persistence.
- Performed user journey analysis across web to provide insights on platform's highest touch points for ML efforts to converge on.
- Analyzed navigation vs search activity to improve our UI, cover all our users' intent to drive north star metrics.
- Built statistical significance framework for experimentation at scale.

Forethought San Francisco, CA
Implementation Engineer July 2020 to Sept. 2020

On the Customer Experience team, leading all technical requirements and touching all aspects of the business: Engineering, Product, Sales and Customer Success.
Implemented: State-of-the-art NLP models to help clients be geniuses at their job
Involved: Data Engineering, Data Science, Machine Learning/Artificial Intelligence, Business Intelligence -- owning whole data pipeline Post-Sale

-Queried MongoDB to create customer business rules.
-Designed AI Training datasets to feed into XLNet and BERT models using Jupyter Python notebooks.
-Analyzed trained models' performance to deploy best automated NLU models for clients.
-Verified live models' predictions were successful via API calls to clients' Salesforce Help Desks.
-Reduced client's SPAM from 64% to less than 1%.
-Helped save client >$20,000 in human labor cost from Customer Support Agents manually labeling tickets.
-Completed data analysis that contributed to signing of >$400,000 deal with Instacart.

Immuno Concepts Sacramento, CA
Quality Control Analyst July 2010 to Apr. 2019

-Built linear regression models to determine whether or not products were drifting from quality.
-Tracked trends and outliers to make manufacturing recommendations to management to create efficiencies and increase profit margins.
-Created product performance reports to drive key business investments for following quarter.

University of California, Davis Davis, CA
Research Associate Jan. 2005 to Dec. 2008

-Through repeated experimentation explored sigma70 subunit architecture to characterize macromolecular complexes involved in transcription of growth-related genes.
-Narrowed down which protein chain substitution in antibody-derived proteins fit best with research aims in pre-targeting radioimmunotherapy for Non-Hodgkin's Lymphoma.

Volunteering

CoronaWhy Apr. 2020 to June 2020
Machine Learning Engineer
Helping to fight against Coronavirus.

CoronaWhy is a globally distributed, volunteer-powered research organisation of 1000+ members. We're using DS and AI to assist the medical community and policy makers answer key questions related to COVID-19. It's supported by Google, Amazon, NASA and other companies.

I am embedded within the Vaccine/Therapeutics Task team, helping the Paper Study Classification group build baseline models to filter papers based on study design.

Education

Springboard, Data Science Career Track Jan. 2017 to Dec. 2017

University of California, Davis Sept. 2003 to Dec. 2007
Genetics Bachelor's of Science