MEHLIAL KAZMI

Linkedin GIT Portfolio

Email : mehlial.hassan99@gmail.com
Mobile : +90 531 430 62 14

Education

- **Georgia Institute of Technology** — Atlanta, USA
  Masters in Computational Data Analytics; GPA: 3.91 — *2022 - 2024*
- **Lahore University of Management Sciences** — Lahore, Pakistan
  Bachelors in Economics and Mathematics, minor in Computer Science; GPA: 3.60 — *2016 - 2020*

Skills

- **Languages**: Python, R, JavaScript (D3), C++, Scala, Stata
- **Databases**: MySQL, PostgreSQL
- **Data Analysis & Machine Learning**: Stan, caret, A/B testing, data.table, dplyr, pandas, matplotlib, bokeh, scikit-learn, numpy, scipy, plotly, ggplot, rpart, networkX, pandas-profiling, gensim, NLP, OpenCV, Shap, PyTorch, Tensorflow, Microsoft Machine Learning Studio, Apache Superset
- **Web Development & Visualization**: Rshiny, Flask, D3, Tableau, DataBricks
- **Cloud Services**: GCP, AWS, Azure ML
- **Big Data & Distributed Computing**: PySpark, Hadoop

Experience

- **Afiniti** — Istanbul, Turkey
  Senior Data Scientist — *2020 - Present*
  - **Revenue Generation for AT&T**: Leveraged Bayesian statistics and Machine learning models to contribute to generating over $3M in incremental monthly revenue for our client, AT&T.
  - **BiasGuard**: Built a Python-based solution using decision trees and LightGBM to ensure data consistency and reduce bias in A/B testing. Optimized models through hyperparameter tuning, achieving 15% improvement in experimental reliability and 20% more accurate performance measurement. Enabled actionable insights by mitigating skews and ensuring robust test group comparisons.
  - **Text-Driven Churn Scores**: Analyzed 2.7M historical customer transcript records to assess churn probabilities. Transformed textual data into embedding spaces using BERT and clustered these embeddings with HDBSCAN. Identified a significant number of customers in the cluster linked to service cancellations. By calculating the Euclidean distance from each customer's embedding to this cluster's centroid, derived a churn score, with shorter distances indicating a higher likelihood of churn.
  - **Data Drift Detection**: Employed caretEnsemble to develop an ML model that effectively identified prediction and feature drift across over 1.5M data points. The findings were crucial in diagnosing the root cause of a 15% decline in model performance.
  - **Advanced Feature Selection**: Utilized scikit-learn's Random Forest to assess feature importance using Gini and Permutation algorithms. Enhanced feature selection with SHAP, leading to a 10% improvement in production models' performance, efficacy, and accuracy.
  - **Optimization Gap Alert**: Developed an automated tool to track and alert on data points excluded from performance ranking models. The solution identifies unmeasured entities and interactions, enabling targeted optimization and improving operational oversight. Enhanced model coverage by ensuring consistent monitoring of gaps in performance data, driving 5% better insights and operational efficiency.
  - **Performance Debugging Dashboard**: Led a team in developing an interactive automated technical dashboard for diagnosing and resolving ML model issues, such as concept drift. This tool reduced model debugging time by 5 hours per week.
  - **Agent Rankings**: Devised agent rankings in a call-center setting using Bayesian statistics and probabilistic programming language tool - Stan. Additionally, built a Deep Learning model with TensorFlow for an alternative ranking methodology.

Academic Projects

- **Hotel Reviews Dashboard**: Conducted an in-depth analysis of a vast hotel reviews dataset comprising 505k observations. Employed the BERTopic natural language processing framework to discern and highlight the primary topics within the reviews. Additionally, developed an interactive Tableau dashboard, enabling the visualization of main topics over time for individual hotels, facilitating proactive decision-making.
- **Detecting Social Media Toxicity**: Developed a robust methodology for detecting toxic comments using the Wikipedia Detox dataset. Conducted comprehensive data preprocessing and employed word2vec to generate word embeddings, serving as feature inputs for various machine learning models such as Logistic Regression, KNN, Naive Bayes, Random Forest, and AdaBoost. Identified the model with superior performance across multiple key metrics.
- **BrainMRI-GPT**: Designed a dialogue-driven tool for brain tumor MRI analysis, utilizing CLIP and a custom lightweight model for precise patch-level tumor localization. Integrated results with Llama-3.2-11b vision model and engineered prompts, allowing intuitive and interactive MRI interpretation.
- **Semantic-Based Recommendation Model**: Leveraged the Amazon Fine Foods review dataset to create a sophisticated recommendation system. Transformed textual data into an embedding space, extracting nuanced semantic meanings through the distilroBERTa model. These embeddings were then utilized as inputs for a Neural Collaborative Filtering (NCF) model, enhancing the accuracy and relevance of the recommendations provided.