Mihir Ojha

• *Email:* mihir.ojhaa@gmail.com • *Portfolio:* https://mihirojha.github.io/ • *Phone:* +1 7783239851

Education

- **Simon Fraser University** June, 2022
  Bachelor of Science - Data Science *Burnaby, BC*
  - Relevant Coursework: Operating Systems, Data Structures, and Algorithms, Statistics, Machine Learning, Networking, Databases

Skills

- **Languages**: Python, R, SQL
- **Machine Learning**: Supervised Learning (Regression, Decision Trees, Random Forest, Gradient Boosting, XGBoost), Unsupervised Learning (k-means Clustering, Dimensionality Reduction, PCA), Neural Networks (ANN, CNN), Deep Learning (Convolutional Neural Networks), Hyperparameter Tuning
- **Frameworks**: Scikit-learn, NLTK, SpaCy, TensorFlow, Keras
- **Data Analysis and Modeling**: Pandas, Numpy, SciPy, Scikit-learn, Data Wrangling, Data Cleansing, Data Preprocessing, EDA
- **Data Visualisation**: Python's Matplotlib, Seaborn, Plotly, Data Analysis Expressions (DAX) in Microsoft Power BI and Tableau
- **Extra**: Git version-control, GitHub, HTML, CSS, Agile, Scrum, Linux, Command Line Interface, Heroku, Railway, Streamlit

Experience

- **Chartered Professional Accountants of British Columbia** Burnaby
  Data Scientist Consultant *Jan 2022– April 2022*
  - Implemented extensive data cleaning and pre-processing techniques on a dataset containing information on more than 50,000 members to improve the quality of the data using Python.
  - Utilized Power BI and ArcGIS to produce visualizations and maps identifying target areas for recruitment and demographic breakdowns of members.
  - Utilized machine learning including various regression algorithms like a ridge, lasso, and Linear to predict the number of members that will join over the next 5 years.
  - Recommended a 5-year plan in resource allocation, recruitment, and data storage thus reducing 50 hours/week and saving 40% in database management costs. Presented recommendations to the Vice President of CPABC.

- **Jain University** Remote
  Research Assistant *August 2022 - December 2022*
  - Conducted extensive data exploration and analysis on an RCSB Protein Data Bank (PDB) dataset, which contains over 15,000 rows and more than 5,000 classes. Utilized advanced data analysis and modeling techniques to improve the dataset's quality and increase the accuracy of models.
  - Implemented data cleaning and pre-processing methods to improve the dataset's quality and increase the accuracy of the models.
  - Developed and applied machine learning models, such as Random Forest, XGBoost, and Neural Networks to predict protein structures and classify proteins based on their properties. Achieved an accuracy of 95% on the test dataset using the Random Forest model. Presented findings and results to the project supervisor.

Projects

- **Telecom Churn Prediction (Classification, EDA, Data cleaning, Data pre-processing, Streamlit, Railway)**:

  Developed a predictive model to analyze customer behavior and predict churn in a telecom company using Python and machine learning techniques. Utilized data pre-processing, feature engineering, and model-tuning techniques to improve performance. Deployed the model as a Streamlit app and hosted on Railway, achieving an accuracy of 95% on the test dataset.

- **Hate Speech Classification via Tweets (NLP. Spacy, NLTK, Deep learning, Sentiment Analysis, Streamlit, Railway)**:

  Developed a machine learning model for classifying hate speech in tweets using NLP techniques and Python. Utilized techniques such as regex, tokenization, lemmatization, and stop word removal to preprocess the data. Achieved an accuracy of 96% on the test dataset by using various algorithms like Naive Bayes, Decision tree, Random Forest, SVM, and deep learning. Deployed the model as a Streamlit app and hosted on Railway

- **Spotify playlist recommendation (Spotipy API, PCA, Dimensionality reduction, K-means clustering)**:

  Developed a music recommendation system using Spotify's API (Spotipy) and Python, utilizing K-Means Clustering to group similar songs and generate personalized playlists. Implemented data pre-processing techniques such as PCA and dimensionality reduction to improve the performance of the clustering algorithm. Provided links to the 11 playlists to the end-users. Achieved a high user satisfaction rate by providing personalized and diverse playlists.