# **Full Name**

[LinkedIn Link](https://www.linkedin.com) | [email-id@xyz.com](mailto:abcd@xyz.com) | +91 0123456789 | [Github-link](https://github.com)

**EDUCATION**

**Fergusson College (Autonomous), Pune**

*Master of Science, Data Science* *(CGPA: 9.83) 2021 - 2023*

* Core coursework in: Python, Databases, Linear Algebra, Optimization, Machine Learning, Neural Networks, Deep Learning.

*Bachelors of Science, Statistics* *(CGPA: 9.63) 2018 - 2021*

* Core Coursework in: Linear Algebra, Multivariable Calculus, Theory of Estimation, Hypothesis Testing, Regression, Design of Experiments, Operations Research, Time Series.

**PROJECTS**

**American Express - Credit Default Classification** [*Github*](https://github.com)

*Numpy, Pandas, Scikit-learn, XGBoost, LightGBM*

* Built a **classification model** to predict the probability that a customer defaults on their credit based on **5.5 million records** and **191 anonymized features**, using the data provided by American Express.
* Engineered new features by taking different aggregations over time. Made predictions using a **Soft Voting Ensemble** of the best performing **XGBoost and LightGBM Classifiers.**

**H&M - Personalized Product Recommendations**

*CuDF, LightGBM*[*Github*](https://github.com)

* Designed a **product recommendation system** to recommend products for customers based on **31.8 million user transactions** using the data provided by H&M.
* Created a custom lightweight **candidate retrieval method** using a combination of candidates that were purchased together in the last week as well as most popular candidates based on age group.
* **Ranked candidates** using a **LightGBM** **model** based on features created using the frequency of purchase and the proportion of customers purchasing the product.

**Jigsaw - Multilingual Toxic Comment Classification** [*Github*](https://github.com)

*Tensorflow, HuggingFace, Transformers*

* Built a **multilingual text classification model** to predict the probability that a comment is toxic based on **4,35,775** **text comments** in **7 different languages**, using the data provided by Google Jigsaw.
* Fine-tuned **BERT-Multilingual-base and XLMRoBERTa** models on the multilingual text.

**Financial Dashboard for Market Intelligence** [*Github*](https://github.com)

*Numpy, Pandas, NLTK, PyTorch, HuggingFace, Transformers* [*Streamlit App*](https://streamlit.io/)

* Built an **end-to-end Financial Dashboard** that displays the Summary, Sentiment Score and important Keywords of the text extracted from the 10-K SEC filings. The dashboard was deployed using Streamlit.
* Each filing had over 34,000 words and was collected using the SEC-API for 12 companies.
* Fine-tuned **RoBERTa, FinBERT and DistilBERT** models for Sentiment Analysis and **T5, DistilPEGASUS and DistilBART** models for Summarization. RAKE NLTK was used for Keyword Extraction.

**OPEN SOURCE EXPERIENCE**

**Scikit-learn | HuggingFace Evaluate (***Authored over 15 merged pull requests)*

* Implemented support for **sparse matrices** as input to **Silhouette Score**.
* Added **multi-class support** to **Average Precision Score** by computing it in a One-Vs-Rest manner for each class.
* Created API to expose **Out of Bag scores** for Gradient Boosting based estimators.
* Improved the description of **Logistic Regression** in the user guide to highlight that it is a **regression model** implemented as a classifier and a special case of GLM.
* Introduced a new section in the user guide explaining **Bernoulli** and **Categorical Distributions** in the context of Generalized Linear Models (GLM).
* Created a section in the documentation showcasing the **integration of Evaluate** with the scikit-learn framework.

**NPTEL****CERTIFICATIONS**

* Deep Learning*Score: 90%*
* Machine Learning *Score: 94%*
* Data Analytics with Python*Score: 94%*
* Python for Data Science*Score: 92%*

**SKILLS & INTERESTS**

Python | SQL | Data Mining | Data Visualization | Natural Language Processing (NLP)