

Vibhor Tyagi

vtyag@illinois.edu[LinkedIn](#)[Portfolio](#)

EDUCATION

- **University of Illinois at Urbana-Champaign** Urbana, IL
Master of Science - Computer Science; Score: 4.0/4.0 Jan 2024-May 2025(Expected)
- **SRM Institute of Science and Technology** Chennai, India
Bachelor of Technology - Computer Science and Engineering; Score: 87.4% July 2017-June 2021

SKILLS SUMMARY

- **Languages / Database Management:** Python, C, C++, Java, SQL, R, MySQL, PostgreSQL, MongoDB, NoSQL, Hive / HiveQL, Spark / PySpark
- **Cloud Platforms / Low Code Tools:** Tableau, Power BI, AWS (S3, EC2, Lambda), Alteryx Designer, Power Automate, UiPath, Power Apps, Version Control (Docker, Git, Github)
- **Python / Machine Learning Frameworks:** Numpy, Pandas, Scikit-learn, XGBoost, LightGBM, CatBoost, H2O.ai, spaCy, NLTK, PyCaret, Hugging Face Transformers, BERT, TensorFlow, PyTorch, Keras, Matplotlib, Seaborn, Plotly, PySpark, Shiny, Streamlit

WORK EXPERIENCE

- **Research Assistant** March 2024 - August 2024
University of Illinois Urbana-Champaign
 - Automated data acquisition pipelines using Selenium and Scrapy for large-scale web scraping, integrating API endpoints to gather diverse data sources on corporate sustainability and carbon footprints.
 - Utilized Pandas, NumPy, and Dask for high-performance data manipulation and preprocessing, handling missing values, outliers, and standardizing feature distributions, improving downstream model accuracy.
 - Conducted in-depth exploratory data analysis using Matplotlib, Seaborn, and Plotly, visualizing trends in environmental regulations and carbon emissions.
 - Leveraged domain-specific feature engineering via Sklearn and Featuretools, implementing advanced feature selection techniques like Boruta and Recursive Feature Elimination (RFE) for optimal model performance
 - Built random forests, gradient boosting (using XGBoost and LightGBM ensemble methods), and logistic regression models to predict corporate non-compliance based on regulatory data.
- **Data Automation Engineer 2** September 2021 - June 2023
PricewaterhouseCoopers
 - Developed and deployed supervised, unsupervised and semi-supervised machine learning models (using FastAPI and MLFlow) for real-time predictions in cloud environments, supporting scalable, low-latency data processing.
 - Built and optimized ARIMA, SARIMA, and Prophet models to forecast financial trends and expenses. Applied LSTM (Long Short-Term Memory Networks) using Keras and TensorFlow for high-precision time series forecasting.
 - Applied K-Means, DBSCAN, and Hierarchical Clustering to uncover hidden patterns and customer segments within datasets, leveraging PyCaret for quick model prototyping.
 - Developed impactful visualizations and optimized SQL queries, reducing database response time and contributing to an 8% increase in client engagement, as evidenced by enhanced click-through rates and user interactions.
- **Undergraduate Research Assistant** Jan 2021 - May 2021
SRM Institute of Science and Technology
 - Spearheaded the implementation of a cutting-edge movie recommendation system leveraging K-means, collaborative filtering, and SVD, resulting in significantly enhanced personalized content delivery.
 - Improved system sparsity and accuracy through innovative use of implicit and explicit feedback mechanisms, coupled with Association Rule Mining.
 - Optimized key metrics including precision, recall, and F-measures, leading to substantial improvements in recommendation quality.
 - Published related paper on IEEE : <https://ieeexplore.ieee.org/document/9395759>

RECENT PROJECTS

- **Automated Detection of Diabetic Retinopathy using Deep Learning**
 - Leveraged Transfer Learning with pre-trained models like ResNet, InceptionV3, and EfficientNet to enhance the performance of the Convolutional Neural Network (CNN) for detecting diabetic retinopathy.
 - Added Explainable AI component by applying SHAP and LIME frameworks to interpret model predictions, providing transparency in decision-making for ophthalmologists and improving trust in automated diagnoses.
- **Advanced Text Analytics for Sentiment and Topic Modeling**
 - Developed a state-of-the-art multilingual document classification system capable of classifying documents into various categories across multiple languages, leveraging transformer-based architectures.
 - Utilized Hugging Face Transformers to implement the XLM-RoBERTa (Cross-lingual Language Model) for multilingual text representation, enabling classification across diverse languages with minimal performance degradation.
 - Fine-tuned it on a multilingual corpus using PyTorch for efficient training and inference.
 - Employed spaCy and Hugging Face Tokenizers for language-specific tokenization, ensuring accurate token representation for multiple languages.
 - Achieved an average F1-score of 92% across 10 different languages on the Jigsaw Multilingual Toxic Comment Classification Dataset