

## SUMMARY

Data Science student with practical experience in building scalable data pipelines and predictive models. Skilled in Python, SQL, and ETL processes, with a strong foundation in data quality principles and big data analytics. Proven ability to extract actionable insights and collaborate effectively in team environments, contributing to innovative data quality improvements.

## EDUCATION

### University of Illinois Urbana-Champaign, Siebel School of Computing and Data Science

**May 2026***Bachelor of Science, Data Science*

- **Coursework:** Foundation in Data Science, Applied Data Science with Python, Linear Algebra, Probability and Statistics for Computer Science, Data Structures in Python, Intro to Machine Learning, Machine Learning in Physics

## EXPERIENCES

### Illinois Data Science Club

**Aug 2023 - Present***Project Manager**Champaign IL*

- Built scalable **data pipelines** to process **25GB** of F1 telemetry and weather data, ensuring **ETL** processes adhered to high data quality standards using **Pandas**, **NumPy**, and **scikit-learn**.
- Trained **ML models** using **RandomForest** and **XGBoost** while incorporating rigorous data quality checks through grid search optimization for hyperparameter tuning to predict race-winning strategies.
- Developed comprehensive **data visualizations** to analyze feature importance, uncover key race-winning factors, and support refined data quality insights for **actionable** decision-making.
- Designed **predictive pipelines** leveraging ensemble learning models and automated model evaluation with **cross-validation** and custom scoring functions, integrating data quality metrics for improved reliability.
- Secured **2nd** place out of 25 teams at IDSC's Data Dive Competition, demonstrating effective **team collaboration** and **commitment** to high data quality standards.

### 2025 Illinois Statistics Datathon

**Mar 2025***Data Scientist**Champaign IL*

- Built an end-to-end data science solution combining **financial forecasting** and **fraud risk analysis** across tokenized and relational datasets, incorporating automated data quality validations to guide strategic credit decisions.
- Developed a chain forecasting model with **linear regression** to **predict** quarterly account spending after performing thorough data quality checks, achieving an RMSE of 1200 despite high variance in user behavior.
- Engineered a **RandomForest** classifier to detect fraud risk with **86%** accuracy, integrating spending predictions and initial data quality assessments to compute fraud-adjusted credit line recommendations.
- Selected as a **finalist** out of **180** teams, recognized by Synchrony Financial judges for technical maturity, business alignment, and clear articulation of **data quality** practices.

## PROJECTS

### NFL Injury Forecast

**Mar 2024**

- Built a custom **web scraper** to collect player injury data across all NFL teams from 2004–2022 directly from the official NFL website.
- Developed a linear regression model to predict whether teams should retain previously injured players based on historical recovery and performance trends across positions and injury types.
- Achieved **60%** accuracy in binary retention prediction, tuning hyperparameters to **balance overfitting** and **generalizability**.
- Utilized advanced **SQL** queries to extract and manipulate **large datasets**, ensuring high **data accuracy** and integrity to support downstream **statistical analysis** and machine learning workflows.

### Candidate Recommendation Engine

**Aug 2025**

- Developed an AI-driven web app that ranks candidates for job descriptions using fine-tuned **SentenceTransformer vector embeddings**, achieving ROC AUC of 0.81 and accuracy of 71%, significantly improving **semantic matching accuracy**.
- **Computed** cosine similarity between job and resume embeddings to efficiently identify the top 5 most relevant candidates in under 10 seconds per query, enhancing recruitment speed.
- **Built** a multi-page PDF resume parser capable of correctly grouping resumes for 100% of tested candidates across 50+ files, ensuring flawless data organization.
- Integrated Google Gemini API to generate automated, personalized candidate summaries, highlighting key skills, gaps, and hiring recommendations, reducing manual review time by **40%** and boosting productivity.
- **Deployed** an interactive Streamlit UI enabling recruiters to filter, rank, and download candidate summaries efficiently, improving hiring workflow and user experience.

## SKILLS

- **Languages/Tools:** Python, SQL, R, Java, C++, HTML, Pandas, NumPy, Tableau, AWS, Excel, Azure Cloud, GCP
- **Techniques:** Data Preprocessing, Data Wrangling, NLP, Big Data Analytics, Data Mining, Data Pipelines, Data Visualization
- **Other:** Analytical Reasoning, Critical Thinking, Team Collaboration, Problem Solving, Ethical AI