# CHAITHRA K . C

(315) 728-0657 • ckoppara@syr.edu • www.linkedin.com/in/chaithra-kc • https://github.com/chaithrakc

## EDUCATION

**SYRACUSE UNIVERSITY, SCHOOL OF INFORMATION STUDIES** **Syracuse, New York**
**Master of Science Applied Data Science** **May 2023**
Coursework: Introduction to Data Science, Applied Machine Learning, Big Data Analytics, Applied Deep Learning, Natural Language Processing, Inferential Statistics, Business Analytics, and Data Analysis & Decision-Making, and Scripting for Data Analysis

**JAIN UNIVERSITY** **Bangalore, India**
**Bachelor of Engineering Computer Science** **Jun 2015**
Relevant coursework: Principles of Programming Languages, Java & J2EE, Data Structures, Algorithms, Database Management Systems, Data Mining, Business Intelligence, Design Patterns, Unix System Programming, and Linux Internals

## TECHNICAL SKILLS

**Key Skills:** Machine Learning, Natural Language Processing, Deep Learning, Data Cleaning, Analysis and Visualization, Hypothesis Testing, Probability and Statistics
**Programming Languages**: Python, R, Java, SQL, PL/SQL, NoSQL
**Python Libraries**: PySpark, Numpy, Pandas, Matplotlib, Seaborn, Scikit-learn, NLTK, Spacy, Keras, TensorFlow, Gensim, Beautiful Soup, PDFMiner
**Databases**: MySQL, Oracle DB, Mongo DB
**Tools & Utilities:** Databricks, Apache Spark, Tableau, Advanced MS Excel, MS Access, Perforce, Git, Maven
**IDEs:** Google Colab, Jupyter Notebook, PyCharm, R-Studio, IntelliJ

## WORK EXPERIENCE

**Data Science Intern, RSG Media, New York** **May 2022 - Aug 2022**

- Developed a workflow using Databricks and Apache Spark to automatically feed historical data as input to XGBoost ML Model by converting existing SQL Queries to PySpark, which was utilized in predicting future audience
- Implemented python module for entity resolution between Viacom's program inventory & Gracenote data, and IMDb & TMDB using NLP pre-processing techniques, TF-IDF word vectorization, and K-Nearest Neighbor ML model
- Built a data pipeline for ingesting 700 million metadata records from IMDb website to AWS S3 bucket using Python AWS SDK Boto3, after which the data underwent various data transformations and validations before being loaded into Databricks Delta Tables
- Aggregated data from several delta tables using SQL and PySpark to generate forecast report that allowed for the comparison of predictions made by various models using MAPE and RMSE error metrics
- Converted SQL-based ETL pipelines source code to PySpark to enable automation

**Senior Software Engineer, Envestnet Yodlee, Bangalore, India** **Mar 2018 - Sep 2020**

- Developed multi-threaded data quality tool in Java to validate holding and transaction data veracity and extrapolate missing financial details based on patterns observed with other users
- Implemented POC for determining suitable database for the project among OLTP (Oracle DB), Mongo DB, and Apache Kudu based on query performance
- Designed Mongo DB Collections, documents with suitable indexing, and configured Spring repositories for application querying
- Implemented Netflix-Ribbon load balancer to achieve fault tolerance and distribution of application traffic
- Created REST API using JAX-RS for sharing reconciliation results with other teams

**Software Engineer, Envestnet Yodlee, Bangalore, India** **Jul 2015 - Feb 2018**

- Built web crawlers using Java Selenium APIs to aggregate users' financial data and personal information from banking websites
- Created PDF parser for retrieving data from bank statements using Python PDFMiner text extraction tool
- Implemented regex validator using Java regular expressions for validating PII
- Reviewed the code submitted by peers and mentored couple of new hires

## ACADEMIC PROJECTS

**Credit Card Default Prediction** **Sep 2022 – Dec 2022**

- This project used three machine learning models (Logistic Regression, Random Forest, and Deep Learning) to analyze credit card default risk without relying on credit scores or credit history
- Dataset consisted of 30,000 credit card users and 26 features
- Random Forest model performed the best, with a precision of 0.80 and a recall of 0.65
- Most important default predictors were the most recent two months' payment status and credit limit
- These models could be helpful for credit card firms, loan lenders, and banks to make educated decisions on creditworthiness

**Sentiment Analysis of COVID News Articles** **Mar 2022 – May 2022**

- The dataset includes news articles related to COVID-19 from UK, India, Japan, and South Korea newspapers**.** On the dataset of 10K news articles that were collected from IEEE data port
- Data cleaning, pre-processing, and exploratory analysis were performed on the dataset
- Textual data was transformed into numeric vectors using word vectorization techniques, such as the Bag of Words (BOW) model, BERT word embedding, and XLNet embedding
- Using Naïve Bayes classifier and BiLSTM neural network, predicted sentiment polarity - Negative, Positive, and Neutral of each news article in the dataset