CHAITHRA K . C

(315) 728-0657 • [ckoppara@syr.edu](mailto:ckoppara@syr.edu%20)  •  [https://www.linkedin.com/in/chaithra-kc/](%20https://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, Data Analysis & Decision-Making, Scripting for Data Analysis, Advanced Bigdata Management.

**GPA: 4.0/4.0**

**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

**CGPA: 9.56/10.0**

# TECHNICAL SKILLS

**Key Skills:** Machine Learning, Natural Language Processing, Deep Learning, Exploratory Data Analysis (EDA), Data Visualization, Hypothesis Testing, Probability and Statistics, RESTful APIs, Data Pipelines, Database Management, Web Scraping

**Programming Languages**: Python, R, Java, SQL, PL/SQL, NoSQL

**Python Libraries**: PySpark, Numpy, Pandas, Matplotlib, Seaborn, Scikit-learn, NLTK, Spacy, Keras, TensorFlow, Gensim, PDFMiner, Selenium, Beautiful Soup

**Databases**: MySQL, Oracle DB, Mongo DB

**Tools & Utilities:** Databricks, Apache Spark, Tableau, Advanced MS Excel, 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

* Created automated workflow using Databricks and Apache Spark to feed historical data into an XGBoost ML model, resulting in improved accuracy of future audience prediction
* Designed and implemented a Python module for entity resolution using NLP pre-processing techniques, TF-IDF word vectorization, and K-Nearest Neighbor ML model, resulting in more accurate matching of Viacom's program inventory with Gracenote data, and IMDb with TMDB
* Built a data pipeline to ingest 700 million metadata records from IMDb website into AWS S3 bucket, which underwent various data transformations and validations before being loaded into Databricks Delta Tables
* Generated forecast report by aggregating data from several Delta tables using SQL and PySpark, allowing for comparison of predictions made by various models using MAPE and RMSE error metrics
* Converted SQL-based ETL pipelines source code to PySpark, enabling automation and improving efficiency of data-processing

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

* Developed multi-threaded data quality tool in Java to validate the holding and transaction data, and to extrapolate missing financial details based on patterns observed in other users
* Conducted a proof of concept to determine the most appropriate database for the project, including OLTP (Oracle DB), Mongo DB, and Apache Kudu, based on query performance
* Designed Mongo DB collections and documents, optimized indexing, and configured Spring repositories for efficient application querying
* Implemented Netflix-Ribbon load balancer to achieve fault tolerance and distribution of application traffic
* Designed and implemented RESTful API using JAX-RS to enable easy integration of reconciliation analysis with other systems

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

* Developed web scraping tools with Java Selenium APIs to aggregate financial and personal data from banking sites
* Designed a PDF parser for extracting bank statement data using Python’s PDFMiner (text extraction tool)
* Implemented regex validator with Java regular expressions for validating Personally Identifiable Information (PII)
* Performed code reviews, involved in the design discussions and mentored new joiners

# 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
* BiLSTM model achieved an accuracy of 64%, whereas Naive Bayes algorithm had an accuracy of 78%