# CHIRAG SUBRAMANIAN

(224) 622-1395 | chirag1810@gmail.com | linkedin.com/in/chiragsubramanian

## SENIOR DATA SCIENTIST

***Leveraging strong experience and expertise in SQL, Python, PySpark, PyTorch, R, Tableau, PowerBI and Microsoft Azure Databricks to manage complex projects and programs***

Innovative Senior Data Scientist holding a Master of Science in Operations Research from Northeastern University, USA, and having 7+ years of full-time work experience in delivering valuable business insights through data analytics and data science. Expert in articulating and translating business questions to both technical and non-technical audiences. Strong experience in Big Data (> 40 billion rows), R, Python, SQL, PySpark, Tableau, PowerBI, Microsoft Azure Databricks, and PyTorch, including leveraging data to articulate the commercial value of data and analytics to senior stakeholders/clients through data visualization and storytelling.

## CORE AREAS OF EXPERTISE

- ❑ Data Mining & Modeling
- ❑ Machine Learning
- ❑ Deep Learning with PyTorch
- ❑ Reinforcement Learning
- ❑ Statistical Distribution Fitting
- ❑ numpy, pandas, scipy, scikit-learn
- ❑ ggplot2, matplotlib, dplyr, data.table
- ❑ Text Mining, NLP
- ❑ R, Python, SQL, Tableau, Spark, PySpark
- ❑ tidyverse, tidyr, tidytext, fitdistrplus
- ❑ Tableau, PowerBI
- ❑ A/B Testing, Hypothesis Testing

## SELECTED CAREER HIGHLIGHTS

- Built a Customer Behavioral Progression propensity model for Walgreens Front of Store in Python using XGBOOST for multi-class classification. Wrote an efficient data processing PySpark pipeline and successfully scored dataset of 200M+ records (all customers) using the trained XGBOOST model. Used insights from the last decision tree in the XGBOOST ensemble to make recommendations to the business.
- Built and deployed a Credit Card Propensity Tool prediction framework (propensity model) for Walgreens Front of Store Customers in Python using an XGBOOST multi-class classification model. Wrote an efficient data processing PySpark pipeline and successfully scored 132M customers using the trained XGBOOST model. The propensity scores from the model are being used to generate new credit card offers for Walgreens customers.
- Engineered complex and efficient SQL queries processing >40B rows of data via Microsoft Azure Databricks, creating a robust Customer Demographics Table and Store Level Insights Dashboard in PowerBI.
- Led the development of Aon Impact Forecasting's U.S. Severe Convective Storm Probabilistic Catastrophe model, spatially simulating their largest ever stochastic event set of 609M rows in R.

## PROFESSIONAL EXPERIENCE

**WALGREENS BOOTS ALLIANCE,** Chicago, IL — May 2022 – January 2024

**Senior Data Scientist**

*I was recruited to the organization to contribute towards statistical machine learning, propensity modeling, statistical modeling, data wrangling and reporting to drive marketing strategies for Walgreens Boots Alliance.*

*TECH STACK - Python, R, SQL, PowerBI, Hadoop, Microsoft Azure Databricks, PySpark*

- Built a Customer Behavioral Progression propensity model for Walgreens Front of Store in Python using XGBOOST for multi-class classification. The model was trained on a dataset of 1.6M records and achieved an average accuracy of 85% across 10 classes on the test set. Wrote an efficient data processing PySpark pipeline and successfully scored dataset of 200M+ records (all customers) using the trained XGBOOST model. Used insights from the last decision tree in the XGBOOST ensemble to make recommendations to the business.
- Built and deployed a Credit Card Propensity Tool prediction framework (propensity model) for Walgreens Front of Store Customers in Python using an XGBOOST multi-class classification model on a dataset of 200K customers with an accuracy of 76% across 4 classes. Wrote an efficient data processing PySpark pipeline and successfully scored 132M customers using the trained XGBOOST model. The propensity scores from the model are being used to generate new credit card offers for Walgreens customers.
- Engineered complex and efficient multi-layered SQL queries processing >40B rows of data via Microsoft Azure Databricks, creating a robust Customer Demographics Table and and visualized it in a Store Level Insights Dashboard using PowerBI.
- Performed K-means clustering in PySpark to segment a dataset of customers into groups based on customer behavior metrics.
- Wrote complex SQL queries to pull data for ad hoc requests from internal business partners.

*Professional experience, continued…*

- Reviewed SQL code written by junior associates and gave them best practices recommendations.
- Mentored and guided junior associates.
- Acted as the lead interviewer for Associate Data Scientist roles in the company.

**HOLLAND AMERICA LINE,** Chicago, IL Oct 2021 – April 2022

**Senior Marketing Analytics Analyst**

*I was recruited to the organization to contribute towards predictive modeling, A/B testing, customer segmentation analytics, statistical modeling and reporting to drive marketing strategies for Holland America Line.*

- Applied the k-means() algorithm in R for optimal customer segmentation.
- Data wrangling in R using dplyr, data.table, tidyverse, tidyr and lubridate packages.
- Data wrangling in Python using numpy, pandas and scipy.
- Writing complex SQL queries fetching data with >8M rows using TOAD for Oracle.
- Creating interactive dashboards and reports using Tableau.
- Automating Excel reports and processes using R.
- Manipulating data using Pivot Tables in Microsoft Excel.

**AMWINS GROUP,** Chicago, IL March 2021 – Aug 2021

**Data Scientist**

*Worked as a Data Scientist at Amwins Specialty Casualty Solutions (ASCS, a division of Amwins Group). Provide predictive modeling, machine learning and distribution fitting expertise to a team of actuaries at Amwins.*

- Fitting statistical distributions to insurance claims data and selecting best-fit distribution. Implementing actuarial research papers and running goodness-of-fit tests comparing fitted data to empirical distribution.
- Performed principal component analysis (PCA) on a matrix of numeric tf_idf vectors calculated from raw text data, reducing 3214 columns to 100 principal components.
- Built a XGBOOST multi-class classification model on top of insurance claims data after PCA and achieved 91% accuracy on test data after 10-fold cross validation and using text mining techniques in R.
- Queried claims data set of 300,000+ rows using MySQL Workbench and fed query results directly into R using the package RMySQL for further analysis.

**AON,** Chicago, IL Sep 2016 – March 2021

**Senior Analyst** **Sep 2020 – March 2021**

*Provide predictive modeling and machine learning expertise utilizing Multiple Linear Regression, Boosted Decision Tree, K-Nearest Neighbors, K-means, Feed Forward Neural Networks, and Multi-Layer Neural Networks in R. Utilize probabilistic modeling and statistical analysis using R and presented models to teammates. Taught management R programming.*

**U.S. Florida Commission Project**

- Developed predictive models in R on datasets with more than 17 million rows. Machine Learning algorithms developed in R – generalized additive models, generalized linear models, neural networks and regression spline.
- Performed statistical goodness of fit t-tests on claims data after aggregating the data to zipcode resolution using summarize() and group_by(), to compare the difference in means of the modeled data with the observed claims data at a 5% significance level.
- Developed more than 1000 graphs and bar charts for different construction types and compared different sets of data after data cleaning and data wrangling. The visualizations I developed in R were important in assessing the accuracy and quality of Impact Forecasting's catastrophe model for the Florida Loss Commission.

**Analyst - Analytics** **Sep 2016 – Sep 2020**

*Provide predictive modeling and machine learning expertise utilizing Multiple Linear Regression, Boosted Decision Tree, K-Nearest Neighbors, K-means, Feed Forward Neural Networks, and Multi-Layer Neural Networks in R. Utilize probabilistic modeling and statistical analysis using R and presented models to teammates. Taught management R programming.*

**U.S. Severe Convective Storm Model (SCS)**

- Spatially simulated the largest dataset in the history of Impact Forecasting's SCS model - 609 million severe convective storm events, and improved model accuracy by developing a revolutionary 5 km by 5 km grid in R.
- Executed loss calculation for 10 US states in the tornado model portion and wrote code in R for data wrangling, spatial simulation, data visualization, probabilistic modeling, looping and error handling.
- Executed a gap statistic analysis using the K-Means algorithm in R to select the optimal number of tornado clusters.

**CHIRAG SUBRAMANIAN** (224) 622-1395| chirag1810@gmail.com

*Professional experience, continued…*

- Increased the size of historical tornado data set from 64,000 events to 90,000+ events, after de-trending; created a comprehensive stochastic set of 26 million tornado events covering the entire United States in R. Successfully created and visualized historical and simulated tornado data in R through histogram, 3D plots, and contour plots.
- Prepared presentation slides and presented the Tornado model at the IF US Quarterly Team Meeting in October 2019. Performed convolution of damage function for different tornado intensities in R. Currently calculating hit probability for 26 million tornado events in the United States, and calculating loss by state for Tornado peril in R.

**U.S. Hurricane**

- Cleaned, sorted hurricane data, preprocessed data, and performed linear interpolation. Wrote code in R using the k-nearest neighbors' algorithm to generate simulations of hurricane track and central pressure, which compared well with historical measurements.
- Wrote code in R for linear interpolation, data cleaning, predictive modeling, simulation and data visualization.

**SMILEYGO**, San Francisco, CA June 2015 – Nov 2015

**Algorithm Manager**

*Developed the SmileyGo search algorithm leading a team of 6 data analyst interns.*

- Developed SmileyGo's Search Algorithm to evaluate and rank 5457 nonprofits using Boosted Decision Tree in R.
- Created informative infographics and reports about education, poverty, homelessness, and other causes.
- Represented SmileyGo with team members at the Impact Challenge of the Business Today International Conference in November 2015.

**Algorithm Lead Intern**

*Researched technologies for developing a machine learning algorithm for the corporate social responsibility industry.*

- Built the first machine learning algorithm for the corporate social responsibility (CSR) industry.

## EDUCATION, CERTIFICATIONS & PROFESSIONAL DEVELOPMENT

GEORGIA INSTITUTE OF TECHNOLOGY, ATLANTA, GA, USA
***Master of Science in Analytics – Specialization in Computational Data Analytics (Expected 12/2024)***

***Relevant Courses:*** *Deep Learning, Reinforcement Learning, Machine Learning, Regression Analysis, Database System Concepts and Design, Computing for Data Analysis, Introduction to Analytics Modeling, Data and Visual Analytics, Deterministic Operations Research, Business Fundamentals for Analytics, Data Analytics in Business, Applied Analytics Practicum*

NORTHEASTERN UNIVERSITY, BOSTON, MA, USA
***Master of Science in Operations Research (Graduated 08/2016)***

***Relevant Courses:*** *Deterministic Operations Research, Applied Probability and Statistics, Probabilistic Operations Research, Data Mining in Engineering, Statistical Data Mining, Optimization and Complexity, Master's Thesis*

STANFORD UNIVERSITY, STANFORD, CA, USA
***Summer School: Statistics Courses (Earned Credit 08/2015)***

MANIPAL INSTITUTE OF TECHNOLOGY, MANIPAL, INDIA
***Bachelor of Engineering in Mechanical Engineering (Graduated 05/2014)***