# SIMI SUDHAKARAN

Data Scientist

## PROFILE

Recent university graduate in Data Science with 5 years of Software Development experience. Skilled in Python, Machine Learning, SQL, Data Analysis, Quantitative Analytics, Predictive Modeling, Streamlit, Flask, Data Visualization using Python/ R/ MS Excel/ Tableau, NLP (Natural Language Processing) and Software development. I witnessed the transformation of the world from a knowledge economy to information to now Big Data and Artificial Intelligence. I believe this is a small beginning to a much larger and grand transformation that will literally change what it means to be a human. Today's world will be reshaped by Data Science. I want to be part of this new phase of computing. It is only natural that I would pursue a master's degree in Data Science for my growing interest in solving problems by transforming data into useful insights and visualizations.

## CONTACT


**PHONE:** +1 (908)636-8721

**LINKEDIN:** @simi-s

**EMAIL:**
simi.sudhakaran@outlook.com

**WEBSITE:**
https://www.kaggle.com/simithewhiz

https://github.com/simi-s


## EDUCATION

**Saint Peter's University, NJ**
Feb 2020 – Feb 2022

- Master of Science in Data Science (minor Big Data Analytics
- Cumulative **GPA 4.0**
- Class valedictorian
- During my masters, I also worked in a Graduate Assistant project, during which we had to do aspect-based sentiment analytics and provide them inputs about various products based on the customer ratings.

## SKILLS

**Data Science:** Machine Learning, Data and Quantitative Analysis, Data Mining, Predictive Modeling, Text Mining, NLP, Topic Modeling with Latent Dirichlet Allocation (LDA), Sentiment Analytics, Aspect based Sentiment Analytics (ABSA), Statistical Analysis, Hypothesis Testing
**Languages:** Python, R, Java
**Database:** Oracle, MySQL, PostgreSQL
**IDE:** VS Code, Jupyter Notebook, PyCharm, Eclipse
**Cloud:** AWS (S3, Redshift), GCP
**Data Processing:** Hive, HDFS, Apache Spark (Databricks)
**Known Concepts:** GIT, HTML5, CSS, Apache Tomcat, OS (Linux, Windows, Mac)
**Data Visualization & Reporting:** Tableau, Python (Plotly Dash, Matplotlib, Seaborn, Viola), MS Excel
**Web/API Framework:** Streamlit, Flask, Viola

## WORK EXPERIENCE

**Infosys Limited**
Senior Systems Engineer

- Built UI, Automation and Data driven applications using Java, Swing, Oracle, SQL. Development of Cron jobs for executing scheduled tasks. Deployment of applications on Linux.
- Integrated, evaluated, fixed bugs, and provided production support.
- Extensively involved with all the phases of the application life cycle.
- Ensure the best possible performance, quality, and responsiveness of the applications.
- Collaborated primarily with client on creating their data aggregation and document generation systems.
- Got recognized with **INFOSYS BRAVO AWARD** for exemplary work in an automation project.
- Was also awarded **INFOSYS STAR OF THE WEEK** multiple times for out of box thinking and exceptional performance.

# PROJECTS

Graduate Assistant Project

- A project with actual business partners, wherein we did Topic Modelling (using Latent Dirichlet Allocation LDA), aspect-based sentiment analytics (ABSA) to provide the client input on how the various products are performing in different departments according to the customer reviews. We signed a non-disclosure agreement; hence the project details could not be disclosed.

Social Analytics Driven Health Web Application (Capstone Project with business partner)

- In this project, we leveraged social media data gathered from Reddit for the purpose of tracking the prevalence of public interests and understanding public reaction towards identifying high priority conditions and/or disease states. For this project we **Web Scraped** the data from reddit by creating our own API using **Flask** which internally used **Pushshift API**. We did a high-level Text Preprocessing using **NLTK library**, performed exploratory data analysis by creating a Time-Series Graph with **Regression Analysis**, **Wordcloud**, interactive **Knowledge Graph** for relational Taxonomy, performed Topic modelling using **Latent Dirichlet Allocation**, **Sentiment Analytics** using VADER, K-Means Clustering on Sentiment subjectivity and polarity. Thereafter created a Web Application using **Streamlit** and deployed it to **AWS Lamda**.
- This application provides potential public health application for early detection of diseases and means of information distribution for proactive response.

Spotify - Find correlation patterns in Songs – PCA & Machine Learning

- In this project we are trying to find correlation pattern of a song feature with other features. For example: can energy of a song and loudness be correlated? We have analyzed approximately 1million records with around 14 sets of features.
- We have done dimensionality reduction using principal component analysis by standardizing all features and reducing a complex dataset to lower dimension to reveal hidden, simplified structure that often underlie it.
- After the PCA and tuning of the dataset we used various machine learning regression models such as Linear Regression, Random Forest Regression and Decision Tree Regression to train the model and calculate the accuracy score on the test dataset.

GitHub Link: https://github.com/simi-s/spotify-pca-and-machine-learning

IMDB- 100-year Movies Trends – Data Visualization

- The goal of this project was to analyze the Hollywood movie industry over the past 100+years using the IMDB dataset and detect trends
- Performed Data wrangling and exploratory data analysis of the dataset.
- Also, created a Viola dashboard with the EDA charts and graphs.

GitHub Link: https://github.com/simi-s/IMDB-Data-Visualization

Estimate Medical Equipment required for Heart patients during Covid19 Pandemic in USA

- The main purpose of this project is to filter and analyze the estimated percentage of prevalence of heart disease in USA which will in turn help in estimating the total cardiac devices required in a given region.
- The dataset used for this project consisted of data from all the states in USA from the timeframe 2011 to 2020. The 2021 data was predicted using the Linear regression model.
- We clustered the state wise data using K-Mean clustering, so that we can prioritize the states which are at a higher risk and therefore require more medical equipment relative to other states.

GitHub Link: https://github.com/simi-s/Estimate-Medical-Equipment-required-for-Heart-patients-in-Covid19

**Covid19 vs. Flu Hypothesis**

- This project was done on May 2020 when the Covid19 virus was relatively new and there was no vaccine available to fight the virus. At that time there was a debate going on that is COVID-19 more infectious than Influenza and has a higher death rate and hospitalization?
- We focused on comparing the death rate of Covid-19 and the Influenza on focus on rejecting the null hypothesis that the "Covid-19 is just like any common flu".
- Null Hypothesis (H0): Covid19 is not any more dangerous than the normal flu (We will compare the death counts) Covid19 = Influenza
- Alternate Hypothesis (Ha): Covid19 is more dangerous that the common influenza
- Got a department level appreciation for this project!

GitHub Link: https://github.com/simi-s/Covid19-vs-Flu-hypothesis

**Have vaccinations brought down Covid19 deaths? - Machine learning classification project**

- We used the Covid19 dataset to classify and prove whether the vaccinations have brought down the Covid19 death rate or not. We used the Logistic regression classifier, K-Nearest Neighbor classifier, Kernel Support Vector Machine classifier, Decision tree classifier, and Random Forest classifier to calculate the accuracy score. Also, performed an ensemble technique (AdaBoost) to get a better predictive model by combing all the above models.

**Covid19 impact on mobility in various sectors of the world - Interactive exploratory data analytics using Tableau**

- Used a combination of 3 datasets to perform interactive exploratory data analytics on the Covid19 dataset and how it impacted mobility in various sectors of the world. We created Geomaps, bar graphs, line graphs, pie charts, time-series and also performed a forecasted data for the next 3 months on the mobility dataset. In the end, I created an impressive interactive dashboard with infographics analytics. Also, wrote a story about how covid19 has impacted the economy and daily life.

**Extract, Transform and Load project using Python and AWS**

- The main goal of this project was to perform ETL on a dataset. Extract data from Postgres to python. Transform the data to CSV on python and upload the data to AWS S3 bucket. Lastly copy the data to the Redshift database and perform an analytical query on the dataset

**Statistical calculator**

- This was a simple python project to model a descriptive statistical calculator. It will calculate count, percent, frequency, mean, median, mode, range, variance, standard deviation, percentile rank, quartile rank and give the five-point summary of an input dataset or manual input.

GitHub Link: https://github.com/simi-s/statistical-calulator