



# SHOUVIK SHARMA

 [ssharma25@hawk.iit.edu](mailto:ssharma25@hawk.iit.edu)

 +1(312)-459-2008

 400 E 33<sup>rd</sup> Street, Apt 515, Chicago, IL-60616

 [medium.com/@shouvik19](https://medium.com/@shouvik19)

 [github.com/shouvik19](https://github.com/shouvik19)

 [linkedin.com/in/shouvik-sharma19](https://linkedin.com/in/shouvik-sharma19)

## EDUCATION

**Illinois Institute of Technology**  
Chicago • 2019 - 2021

**Master of Data Science**  
• GPA - 3.88/4.0

**NMIMS University**  
Mumbai • 2016 - 2018

**Master of Statistics**  
• GPA - 3.35/4.0

## SKILLS

**Data Science | Analytics:** Linear Regression, Multiple Regression, Logistic Regression, Naïve-Bayes, KNN, Time Series Analysis, AdaBoost, Ensemble classifier, ML Techniques, K-Means Clustering, Hierarchical Clustering, SAS Enterprise Miner, SAS Enterprise Guide, SPSS

**Languages:** Python, R, Spark, Hive, Pig, MapReduce, other Big data technologies

**Linear Algebra | Statistics:** Z-test, ANOVA, Chi-square test, Mathematics

**Certifications:** SAS Certified Base Programmer for SAS 9 in Mar 2017, SAS Certified Predictive Modeler Using SAS Enterprise Miner 14 in Apr 2018, Practical Machine Learning in Dec 2018 from John Hopkins University, Machine Learning Specialization in Feb 2019 from University of Washington, Snowflake Pro Certification September 2020

**Deep Learning:** Convolution Neural Network, Recurrent Neural Network, Long Short-Term Memory Network

**Cloud:** AWS Lambda, AWS S3, AWS EC2, AWS CLI, Kafka, Redshift, AWS Sage Maker, Azure Data Factory, AWS DynamoDB, AWS CLI, AWS VPC, AWS ELB, AWS EBS, AWS AMI

**Databases:** SQL Server, Snowflake, PostgreSQL, MSSQL, Microsoft Visual Studio, Visual Basic, Pentaho

**Data Visualization:** Tableau, R shiny, Power BI

**Presentation:** Prezi Professional (3D PPT)

## ACADEMIC PROJECTS

- **Stack Overflow Data Analysis (October 2019 -December 2019)** (Spark, Hive, Spark, Big Data Technologies)  
Analyzed insights about questions posted on stack overflow by extracting large data sets using Google's big query data warehouse ; discovered top spammers, expert users, and most valuable customers users by leveraging big data technologies such as Apache Hive, Apache Pig and Apache Sparks
- **Recommendation System using Yelp (January 2020 - March 2020)** (Spark, Hive, Spark, Big Data Technologies)  
Built a personalized restaurant recommender web app using the Yelp dataset of restaurants by testing models like Pure Collaborative, Approximate Nearest Neighbour, K-NN, Naive Bayes and Hybrid Matrix Factorization on different hyperparameters which were tuned using the python library scikit optimizer
- **Image Mating using CelebAMask-HQ (June 2019 - July 2019)** (Python, Jupyter Notebook UNET Architecture)  
Conducted Image Matting using the U-Net architecture of the Convolved Neural Networks on the open-source Celeb-Mask dataset with an IOU Score of 92%
- **Inventory Optimization problem on Kaggle (January 2019 - February 2019)** (R Studio)  
Forecasted the demand for LED televisions using Holt-Winter's Smoothing method with MAPE of 20.760
- **Book Recommendations from Charles Darwin (July 2020 - August 2020)** (Python)  
Designed a book recommendation system based on the content utilizing the open-source Charles Darwin's bibliography using the data mining techniques
- **ASL Recognition with Deep Learning (July 2020 - August 2020)** (Python)  
Created a convolutional neural network to classify images of American Sign Language (ASL) letters

## WORK EXPERIENCE

### LabelMaster - Data Scientist (Practicum Student)

Chicago, Illinois • 08/2020 - Present (ML, data visualization, Data Engineering)

- Explored relationship between sales data and 9 freight market data, each with over 200 input attributes, in Google Cloud AutoML
- Visualized correlation between sales and external factors by scatterplot with linear fit, heatmap, and polynomial fit line
- Predicted dept sales using four machine learning algorithms in Google Cloud AutoML, and found random forest have the best performance with percentage error of 1.7% and R square of 90%
- Built user interface dashboard for presenting customized correlation visualization and model prediction through Tableau

### Daten Solutions Inc. -Data Scientist Intern

Schaumburg, Illinois • 05/2020 - Present

- Developed data migration pipeline from SQL Server to Snowflake, and performed dimensional modeling on the migrated data
- Automated ETL processes using Prefect (Python), making it easier to wrangle data sets and reducing time by as much as 40% by performing large-scale data conversions, and transferring BAAN data into standardized formats for integration into Snowflake
- Developed statistical models like ARIMA using statsmodels package in Jupyter Notebook, the model achieved an overall accuracy of MAPE 5.96%

### Cartesian Consulting Inc. - Data Analyst

Mumbai, India 04/2018 -07/2019

- Determined trend for improving customer retention and reducing churn rate using logistic regression, led to a two-fold improvement in the campaign response
- Executed geography-wise analysis by creating customer one view and customer profiling, and translated analysis into business terms and actionable guidance
- Identified the 'Most Valuable Customer' by leveraging the customer data and deploying Random Forest algorithm with True positive rate of 81%, this led to better customer targeting and improve yearly top-line revenue by 13 %

### Greeksoft Technologies Pvt. Ltd. - Strategy and Analytics Intern

Mumbai, India 09/2017 - 12/2017

- Built an RNN Neural Network model for Live positional trading using Keras package in python where outputs supplemented Bull Spread Strategy in Options Trading, the developed model architecture was backtested for the period from year 2012 to year 2017 where it achieved correct market prediction in 71 % of the days ; this forecasting architecture is utilized for live trading

### Tata Capital Financial Services Ltd. - Associate Analyst

Mumbai, India 07/2015 - 07/2016

- Drove acquisition channel of used car and two-wheeler dealerships, by building customer scorecard after analyzing different parameters affecting the repaying capacity
- Led a team of 3 to construct customer risk assessment by analyzing financial reports and client credit history, which led to a multi-fold increase in corporate lending for two-wheeler and used cars segment, with 0% NPA cases reported over the course of 10 months