**Shouvik Sharma**

400 E 33rd Street, Apt 515, Chicago, IL-60616|Phone: 3124592008 | [shouvik19@gmail.com](mailto:shouvik19@gmail.com) | [Linkedin](https://www.linkedin.com/in/shouvik-sharma19/) | [Github](https://github.com/shouvik19) | [Medium](https://medium.com/@shouvik19)

**SUMMARY**

Over 3 years of comprehensive work experience in Data Engineering, Marketing Analytics and Business Intelligence in banking and retail domains. Ability to solve complex business problems using ETL, Data Mining, Machine Learning & Data Warehousing concepts.

**EDUCATION**

* MS in Data Science, Illinois Institute of Technology, **GPA: 3.8** ***(Aug 2019 – May 2021)***
* MS in Statistics, NMIMS University, **GPA: 3.35** ***(Jul 2016 - Apr 2018)***
* Certifications**:** [Snowflake Pro Certification](https://www.youracclaim.com/badges/f03d4251-13bd-4fd0-9f0d-45ff17bd718f), SAS Certified Base Programmer for SAS 9, SAS Certified Predictive Modeler

**WORK EXPERIENCE**

***Data Engineer at Daten Solutions Inc., Chicago: (May 2020 - Present)***

* Developed and automated **data migration pipeline** from SQL Server to Snowflake and performed **dimensional modeling** on the migrated data, further created **data dictionary** for the technical audience.
* 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.
* Created **Tableau** dashboards to explain variation in success **Metrics** and **Time Series Analysis** tohigher management.
* Automated reporting process using **Excel VBA (Macros)** and **MySQL** maintaining accuracy and saving **~ 75%** of time.

***Data Engineer – Practicum Student at Labelmaster, Chicago: (May 2020 – Dec 2020)***

* Involved in designing data pipelines, databases, data marts, E-R model for **OLTP** and multi-dimensional model for **OLAP**.
* Optimized complex **SQL** scripts for quality checking of projects and populating output tables for deployment using **SSIS**.
* Automated hourly status report saving **10 man-hours/week**, thus decreasing response time for fixes and campaign failures.
* Achieved an accuracy of **MAPE 8%** approx. on price forecasting using Deep Learning algorithms like **LSTM** and **RNN**, further created dashboards for presenting the forecasted values to the higher management.

***Data Engineer at Cartesian Consulting: (Apr 2018- Jul 2019*)**

* Developed pipelines for **ETL** (Extract, Transform, Load) using **MySQL**, **Python**, **Airflow** and **AWS S3** foracquiringaPOCproject.
* Extracted data from streaming pipelines using **Flume** and **Kafka** and processed using **Spark** Structured Streaming.
* Predicted sales by **time series forecasting** in **Python** using **neural networks, ARIMAX** and **Prophet** for inventory management by eliminating understocking and reducing overstocking by 56%.
* Applied **K-means clustering** in **Python** for **segmentation** of customers, comparing it with **RFM** (Recency, Frequency and Monetary Value) analysis for improved campaign targeting.
* 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 improving yearly topline revenue by 13 %
* Developed **dimensional** **data** **models** and **data** **warehouse** adhering to integrity and **normalization** rules to support campaign **data** **mart** and customer one view for marketing campaigns. Wrote **complex** **SQL** queries (multiple joins, CTE’s, subqueries).
* Generated visualizations using **Tableau** toanalyze marketing **metrics** for making recommendations and supply chain analysis.

***Data Engineer Intern at Greeksoft Technologies Pvt. Ltd.: (Sept 2017 - Dec 2017*)**

* Identified probable customer churn using **Classification Models** in **Python** like **Decision Trees** and achieved a recall of 84%.
* Worked with the **Apache** **Spark** Framework for customer analytics using **Spark** **SQL** queries on large scale datasets for developing flawless **CRM** (customer relationship management) campaigns and deployed them through multiple channels.
* 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.

***Data Engineer at Tata Capital Financial Services Ltd.:* (*Jul 2015- Jul 2016*)**

* Built **KPIs** and **Regression** models to predict **customer life-time value**, enhance propensity and scoring attributes.
* Accurately extracted insights and created dashboards using **Tableau,** **Excel VBA (Macros)**, **pivot** **tables** and **slicers**.
* Formulated ad-hoc reports based on requirements gathered from various stake holders using **JIRA** to provide solutions.

**PROJECTS**

***Stack Overflow Data Analysis Model (Language/Tools- Python, Jupyter Notebook, Spark, Hive, PySpark, Pig):***

* 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 ([git link](https://github.com/rahulmnair1997/StackOverflow-Data-Analysis))

***Recommendation System using Yelp (Language/Tools- Python, Jupyter Notebook:***

* 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 ([git link](https://github.com/shouvik19/Restaurant-Recommendation-System-using-Yelp-Dataset))

**SKILLS**

* ***Programming:*** SQL, Python, R, SAS, Pyspark, HTML, C#, Excel VBA (Macros), Talend.
* ***Big Data Ecosystem***: Spark, Hadoop, MapReduce, Hive, Pig, Kafka, Flume.
* ***Cloud Technologies***: AWS (S3, EC2, Lambda, Athena, RDS, Redshift, EMR).
* ***Tools:*** Tableau, Power BI, Azure ML, RStudio, Jupyter Notebook, SAS E-Miner, SAS CI, IBM-Unica, SSIS, MS Office, JIRA.
* ***Libraries*:** Numpy, Pandas, Matplotlib, Seaborn, Scikit-Learn, Keras, Nltk, Gensim, Scipy, Beautiful Soup.
* ***Database Concepts:*** Data Services, Data Dictionary, Tool Standards, Data Lineage, User Training