**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 with attributes like analytical thinking and ability to articulate complex situations. Ability to solve complex business problems using ETL, Data Mining, Machine Learning & Data Warehousing concepts.

**LEADERSHIP**

Head of the Sports Department (ISA) – NMIMS, Mumbai, India **(June 2016 – Apr 2018)**

• Led a team of 6 volunteers. Coordinated various workshops on Sports Event for 50+ students.

**EDUCATION**

* **MS in Data Science**, Illinois Institute of Technology, **GPA: 3.8** **(Aug 2019 - May 2021)**

**Related Courses**: Machine Learning, Big Data Technologies, Applied Statistics, Statistical Learning, Database Management,

Data Preparation and Analysis, Introduction to Algorithm, Data Science Practicum.

* **MS in Statistics**, NMIMS University, **GPA: 3.35 (Jul 2016 - Apr 2018)**

**Related Courses:** Regression Analysis, Estimation, Testing of Hypothesis, Distribution Theory, Linear Algebra and Numerical Methods, Parametric Inference estimation, Probability Theory, Linear Models

* 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

**SKILLS**

* **Programming:** SQL, Python, R, SAS, Pyspark, HTML, C#, Excel VBA (Macros), Talend, Agile Methodology, PostgreSQL, MySQL.
* **Big Data Ecosystem**: Spark, Hadoop, MapReduce, Hive, Pig, Kafka, Flume, Hbase, Microsoft Azure.
* **Cloud Technologies**: AWS (S3, EC2, Lambda, Athena, RDS, Redshift, EMR), NoSQL, Cassandra, MongoDB, Kubernetes, Snowflake, CircleCI, Airflow, Prefect.
* **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.
* **Datasets:** HTTP, HTML, XML, JSON

**WORK EXPERIENCE**

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

* Developed and automated **data migration pipeline** from SQL Server to Snowflake using **SnowSQL** and **SnowPipe**, 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 **PowerBI** 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, maintained version control Git, Mercurial, SVN.

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

* Involved in designing databases, data marts, E-R model for **OLTP** and multi-dimensional model for **OLAP** using **SnowSQL**.
* Optimized complex **SQL** scripts for quality checking of projects and populating output tables for deployment using **Azure Pipelines**.
* 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** using **Snowflake**, **Python**, **Azure DevOps** 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.
* 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.

**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, within the required deadline.

**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 **GCP’s big query** data warehouse by leveraging big data technologies such as **Apache Hive**, **Apache Pig** and **Apache Spark** ([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** with an **AUC** of 0.81 ([git link](https://github.com/shouvik19/Restaurant-Recommendation-System-using-Yelp-Dataset))

**Electronic Vendor Database: (Language/Tools - MySQL, Java 8, HTML, CSS, Bootstrap):**

* Constructed the ER Model and translated into Relational Schema implemented as SQL script.