Shouvik Sharma

400 E 33rd Street, Apt 515, Chicago, IL-60616|Phone: 3124592008 | shouvik19@gmail.com | Linkedin | Github | Medium

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.

LEADERSHIP

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

• Led a team of 10 volunteers. Organized various workshops on Fitness and scope of data science in analytics 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, 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, Talend.
- **Cloud Technologies**: AWS (S3, EC2, Lambda, Athena, RDS, Redshift, EMR, Kinesis), NoSQL, Cassandra, MongoDB, Kubernetes, Snowflake, CircleCI, Airflow, Prefect, Workday, Peoplesoft.
- **Tools:** Tableau, Power BI, Azure ML, RStudio, Jupyter Notebook, SAS E-Miner, SAS CI, IBM-Unica, SSIS, MS Office, JIRA, Alteryx.
- **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 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** to higher management.
- Automated reporting process using **Excel VBA (Macros)** and **MySQL** maintaining accuracy and saving **~ 75%** of time, maintained version control Git, Mercurial, SVN.
- Created Talend ETL jobs to receive attachment files from pop e-mail using **tPop**, **tFileList** and **tFileInputMail** and then loaded data from attachments into database and achieved the files.

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

- Involved in designing 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** for acquiring a POC project.
- 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** to analyze 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 with an accuracy of 71%.
- Extracted required discrete & continuous data, assessed data to remove outliers & impact of each measured variable through hypothesis testing and correlation analysis in R and Python.

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.
- Developed an **ETL** pipeline for loading and wrangling data from **SSIS** to **IBM UNICA** for daily campaign workflows using **data warehousing** concepts.
- Pre-processed structured and unstructured data from different sources (**RDBMS**, **Hadoop**) using **Hive, SQL, Sqoop**.
- Automated SFTP process by exchanging SSH keys between UNIX servers. Worked Extensively on **Talend Admin Console** and Schedule Jobs in **Job Conductor**.
- Involved in production n deployment activities, creation of the deployment guide for migration of the code to production, also prepared production run books.
- Self-Starter and Team Player with excellent communication, organizational and interpersonal skills with the ability to grasp things quickly.

Data Engineer at Nielsen*:* **(Apr 2017- Jul 2017)**

- Worked as Data Engineer Intern to automate sample design processes using **R software**.
- Created Implicit, local, and global Context variables in the job. Worked on **Talend Administration Console** (TAC) for scheduling jobs and adding users.
- Assisted in designing and development of technical architecture for sample design process.
- Reduced time required to complete these processes by 25%, thereby helping management to make important decisions faster
- Propose potential research-on-research tests to improve current Nielsen methodologies and improve response and compliance.
- Dive in, and work with our data science team to develop new data-centric products involving new and innovative algorithms.
- Classification of store types based on store attributes using **Random Forest** algorithm in **PySpark** which resulted in better surveying and data collection.

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 BigQuery** data warehouse by leveraging big data technologies such as **Apache Hive**, **Apache Pig** and **Apache Spark** (git link)

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)

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

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