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 Science, Marketing Analytics and Business Intelligence in banking, retail, and supply chain domains. Ability to solve complex business problems using ETL, Data Warehousing, Machine Learning and Exploratory Data Analysis by working independently, and designing analytical solutions.

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), Regex, NLP, Adobe Analytics, GitLab, Amplitude.
- **Big Data Ecosystem**: NoSQL databases, Spark, Hadoop, MapReduce, Hive, Pig, Kafka, Flume.
- **Cloud Technologies**: AWS (S3, EC2, Lambda, Athena, RDS, Redshift, EMR), MATLAB, GCP, EKS, ECS, Glue, Apache Impala.
- **Tools*:*** Tableau, Power BI, Powerpoint, RStudio, Jupyter, SAS E-Miner, SPSS, SSIS, MS Office, JIRA, Spotfire, Databricks, Looker.
- **Libraries:** Numpy, Pandas, Matplotlib, Seaborn, Scikit-Learn, Keras, Nltk, Gensim, Scipy, Beautiful Soup, Tensorflow.
- **Statistics:** Confidence Interval, Regression, Time Series, Cluster, Factor Analysis
- **Scripting:** Linux, Unix

WORK EXPERIENCE

Data Scientist 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 using version control in **GitHub** after breaking down strategic problems.
- Performed customer segmentation using k-mean clustering in **AWS Sagemaker**, further analyzed data to provide subject matter insights and recommended cluster-wise products using **apriori algorithm** which ultimately improved the top-line revenue by 4%.
- Created ad-hoc reports and tableau dashboards to explain variation in success Metrics and **Time Series Analysis**.
- Developed statistical models like **ARIMA** using statsmodels package in Jupyter Notebook, the model achieved an overall accuracy of **MAPE** 5.96%.
- Created sales forecasting predictive model using **Dataiku DSS** using code recipes after extracting data from Snowflake, further automated the pipeline using **scenario** feature of Dataiku DSS, and lastly utilized **API Deployer Node** for deployment.

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

- Predicted department-wise sales based on seasonal and external factors, by working with business stakeholders.
- Implemented Statistical methods like SARIMAX, VAR along with some hypothesis testing as well as Machine Learning (Deep Learning) Time-Series techniques to large sales data.
- Achieved an accuracy of MAPE 8% approx. on price forecasting using Deep Learning algorithms like **LSTM** and **RNN**, to showcase results further created dashboards using **Tableau**.
- Discovered important commodity associated with department sales through feature importance and **ANOVA** analysis.
- Predicted department-wise sales using four machine learning algorithms in Google Cloud **AutoML**, and found random forest has the best performance with percentage error of 1.7% and R square of 90%.
- Extracted data from streaming pipelines using **Flume** and **Kafka** and processed using **Spark** Structured Streaming.

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

- Identified probable customer churn using Predictive Models in Python like Logistic Regression, Decision Trees, Random Forest and achieved a true positive rate (recall) of 84% for target customer retention and acquisition marketing campaigns.
- Predicted sales by time series forecasting using statistical concepts in Python using neural networks, ARIMAX and Prophet for inventory management by eliminating understocking and reducing overstocking by 56%.
- 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 topline revenue by 13 % for a grocery client.
- Generated visualizations using Tableau to analyze marketing metrics for making recommendations and supply chain analysis.
- Performed **marketing mix modeling** and ROI analysis to quantitatively estimate the effectiveness of various marketing elements for one of the trading platforms based in India.
- Performed hypothesis testing to validate whether "Fantasy sports is a game of skill or gamble" using the Chi-Square Test, Linear Regression and paired T-test, the findings successfully published in the Harvard Business Review.
- Performed **failure probability modeling** for energy sectore client to help increase performance, **predict occasional failures** in the functioning and as a result reduce maintenance costs by 15%.

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

- 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 with an accuracy of 71 %.
- Deployed automate end-to-end predictive modeling pipeline using AWS DevOps, where it supported automated daily price forecasting using the LSTM neural network architecture.

Data Science Intern at Nielsen India Inc.: **(Jul 2015- Jul 2016)**

- Worked as Data Science Intern to automate sample design processes using **R software**.
- 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.
- Used Customer 360 tool for computing the extracting a global view of entities.

Data Scientist 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.
- Executed geography-wise analysis by creating customer one view and customer profiling and translated analysis into business terms and actionable guidance.
- Deployed **Feature Selection** using the **Boruta** library in R for determining the most impactful features for predictive modeling.

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 ; discovered top spammers, expert users, and most valuable customers users by leveraging big data technologies such as Apache HiveQL, Apache Pig and Apache Sparks (git link)

Recommendation System using Yelp (Language/Tools- Python, Jupyter Notebook, NumPy, SciPy, pandas, scikit-learn):

- Built a personalized restaurant recommender web app using the Yelp dataset of restaurants by testing models like Pure Collaborative, Approximate Nearest Neighbor, K-NN, Naive Bayes and Hybrid Matrix
- Factorization on different hyperparameters which were tuned using the python library scikit optimizer (git link)

Image Mating using CelebAMask-HQ (Language/Tools- Google Colab, regression):

- Conducted Image Matting using the U-Net architecture of the Convoluted Neural Networks on the opensource Celeb-Mask dataset with an IOU Score of 92%

Inventory Optimization problem on Kaggle (Language/Tools- Google Colab, Tableau, R studio, Adobe Analytics):

- Forecasted the demand for LED televisions using different time-series forecasting methods with Holt-Winter's Smoothing method as the best method with MAPE of 20.760.

Book Recommendations from Charles Darwin (Language/Tools – Spyder, Anaconda):

- Performed nlp techniques like tokenization, stemming, bag-of-words model and tf-idf model for the dataset acquired from project Gutenberg.
- Designed a book recommendation system based on the content utilizing the Charles Darwin's bibliography.

ASL Recognition with Deep Learning (Language/Tools – Spyder, Anaconda):

- Performed one hot encoding using MLLib on the acquired American sign language dataset.
- Created a convolutional neural network to classify images of American Sign Language (ASL) letters

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

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

Word Frequency in Classic Novels (Language/Tools – R Software, NLTK)

- Performed webscrapping using BeautifulSoup and requests libraries in Jupyter Notebook to extract dataset from website Project Gutenberg.
- Further, implemented nltk library in python to analyze unstructured data, and identify the distribution of words.