SHOUVIK SHARMA

shouvik19@gmail.com
+1-312-459-2008
Chicago
linkedin.com/in/shouvik-sharma19
github.com/shouvik19
medium.com/@shouvik19

EDUCATION

MASTER OF DATA SCIENCE
ILLINOIS INSTITUTE OF TECHNOLOGY
August 2019 – Present
Chicago, USA
- **GPA: 3.71**

MASTER OF STATISTICS
NMIMS
July 2016 – April 2018
Mumbai, India
- **GPA: 3.35**

SKILLS

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

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

Programming Languages: Python, R, Spark, Hive, Pig, PySpark

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

Database: SQL Server, Snowflake, PostgreSQL, MSSQL, MYSQL, Microsoft SQL Server, Microsoft Visual Studio, Mongo DB, Cassandra DB

Tools: Tableau, Power BI, Pentaho, MapReduce, Visual Studio, Prefect, SSIS, SSRS, SSAS, SharePoint

Cloud: AWS Lambda, AWS S3, AWS EC2, AWS CLI, Kafka, Redshift, AWS Sage Maker, Apache Kafka

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

EXPERIENCE

DATA SCIENTIST
Daten Solutions Inc.
May 2020 - Present **Chicago, USA**
Daten solutions offer a wide range of consulting services from Analytical solutions to ETL
- Developed data migration pipeline from SQL Server to Snowflake, and performed dimensional modeling on the migrated data
- Improvement performance of existing ETL processes and SQL queries for weekly CRM summary data
- Automated ETL processes using Prefect (Python), making it easier to wrangle data 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
- Led a project to analyze service order demand pattern, and design a demand forecasting model for better resource allocation
- Performed data cleaning, time series transformation, data wrangling, in Jupyter Notebook for data preparation
- Developed statistical models like ARIMA using statsmodels package in Jupyter Notebook, the model achieved an overall accuracy of MAPE 5.96%
- Created interactive dashboard using R Shiny
- Automated the end-to-end model workflow using Azure DevOps and Azure Machine Learning which allows CI/CD architecture

DATA ANALYST
Cartesian Consulting Inc.
April 2018 – July 2019 **Mumbai, INDIA**
- Developed customer insights for one of India's largest grocery chains, to assist their marketing team for improving Customer Retentions, Reducing Churn Rate, Campaign Responses, Lift and Incremental revenue using statistical techniques like RFM methodology, Linear Regression and Logistic Regression.
- Built CLTV & BTYD propensity models using BTYDplus library in R, these models helped to choose best customers for loyalty programs
- Incorporated market basket analysis to improve campaign ROI through cross sell, it improved the topline revenue year-on-year by 3%
- Created & automated various business trend reports & trackers to analyse patterns & movements in business KPIs
- Created interactive dashboards using Qlikview for showcasing key metrics to the senior leadership
- Built recommender-engine to target the Customers with relevant Products using Apache Mahout
- Determined trend for improving customer retention and reducing churn rate using logistic regression, led to a two-fold improvement in the campaign response
- Performed data migration by building ETL pipeline using Pentaho for transfering file from Postgre SQL server to Mariab DB SQL
- 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
- Identified the 'Most Valuable Customer' by deploying Random Forest algorithm with True positive rate of 81%, this led to better customer targeting and improve yearly top-line revenue by 13 %
- Performed marketing mix modeling and ROI analysis to quantitatively estimate the effectiveness of various marketing elements for one of the top fantasy sports platform 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.(link)

EXPERIENCE

STRATEGY AND ANALYTICS INTERN

Greeksoft Technologies Pvt. Ltd.

September 2017 – December 2017 **Mumbai, INDIA**

- Led a price forecasting project Technologies Pvt. Ltd. Mumbai
- Extracted stock price data using NSEpy library which is used to extract historical and real-time data from NSE's website in python.
- Created external variables using technical stock indicators for determining the impact on the closing price
- Performed data cleaning and data manipulation for excluding holidays in the stock market
- Built an RNN Neural Network model for Live positional trading using Keras package with Tensorflow in AWS Sagemaker, where outputs supplemented Bull Spread Strategy in Options Trading, the developed model architecture was backtested for the period from 2012-2017 where it achieved correct market prediction in 71 % of the days ; this forecasting architecture is utilized for live trading
- 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

Nielsen Inc.

May 2017 – July 2017 **Vadodara, INDIA**

- 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

ASSOCIATE ANALYST

Tata Capital Financial Services Ltd.

July 2015 – July 2016 **Mumbai, INDIA**

- Drove acquisition channel of used-car and two-wheeler dealership, by building customer scorecard after analyzing different parameters affecting the repaying capacity
- Leveraged data into innovative features, insights and opportunities through data mining, data cleaning/curating, wrangling, missing values imputation, excluding outliers using R and Python libraries like NumPy, Pandas, SciPy and scikit-learn; Ensured data quality assurance and increased accuracy of machine-learning algorithms such as linear and logistic regression, decision tree, random forest by nearly 13%
- 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
- Spearheaded process of viewing reports by implementing real-time dashboards in PowerBI for operational and exploratory analysis to analyze company's 400+ key performance indicators(KPI) related to People Operations, turnover, recruiting metrics
- Led the design of a 90 node cluster Hadoop ecosystem driven data lake; Wrote complex SQL, HQL and SparkSQL queries to fetch data from multiple sources; Implemented PL/SQL based stored procedures for data ETL and ingestion
- Effective development of row-key driven NoSQL data model in HBase for efficient data retrieval and key performance indicators (KPIs) from reporting standpoint; Improved expensive query efficiency by 65% as compared to RDBMS

DATA SCIENCE INTERN

LabelMaster

Aug 2020 – Present **Chicago, USA**

- 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
- Discovered important commodity associated with department sales through feature importance and ANOVA analysis
- 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

PROJECTS

- **Stack Overflow Data Analysis (October 2019 -December 2019)**

- Analyzed insights about questions posted on stack overflow by extracting data 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, Mongo DB and Apache Spark **(git link)**
- Leveraged big data technologies such as Apache Hive, Apache Pig and Apache Spark for deriving insights about the users.
- Extracted data using Google's big query data warehouse, identified top spammers, expert users, and most valuable customers by using data mining tools like Apache Pig and Apache Hive.
- Built tag prediction model for predicting the tags for a stack overflow post using natural language processing and random forest classifier, the predictive model achieved an accuracy of 72.3 %

- **Recommendation System using Yelp (January 2020 – March 2020**)

- 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)**
- Implemented Natural Language Processing (NLP) based text mining model like Logistic Regression, SVM, K-NN using count vectorizer, n-grams, tokenizer, wordnets from nltk and spacy package in Python to analyze sentiments of unstructured chat transcripts and feedbacks and interpreted models using confusion matrix and ROC curve; Classified transcripts with 92% accuracy resulting better understanding of customer content

- **Image Mating using CelebAMask-HQ (June 2019 – July 2019)**

- In this project we tackle on the problem of background removal through image matting. It consists of predicting the foreground of an image or a video frame.
- Conducted Image Matting using the U-Net architecture of the Convoluted Neural Networks on the open-source Celeb-Mask dataset with an IOU Score of 92% in Spyder Notebook **(git link)**

- **Inventory Optimization problem on Kaggle (January 2019 – February 2019)**

- Forecasted the demand for LED televisions using Holt-Winter's Smoothing method, Facebook Prophet and simple exponential smoothing
- Facebook prophet performed best optimization with MAPE of 20.760 using R software **(git link)**

- **Book Recommendations from Charles Darwin (July 2020 – August 2020)**

- 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 **(git link)**

- **ASL Recognition with Deep Learning (July 2020 – August 2020)**

- 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 **(git link)**

- **Word Frequency in Classic Novels (June 2020 – August 2020)**

- 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