GAURAV GUPTA

Data Scientist (Senior Assistance Manager)

Email: gauravgupta1242@gmail.com
LinkedIn: www.linkedin.com/in/gauravgupta1242
Git Hub: https://github.com/gauravgupta1242
Kaggle: www.kaggle.com/gauravgupta1242
Profile: https://gauravgupta1242.github.io
Contact: 7380835525

PROFILE SUMMARY:

Experienced Data Scientist with 5.5 years in the banking sector, specializing in predictive modeling, model monitoring, and fair lending practices. Proven ability to deliver data-driven solutions and ensure regulatory compliance. Highly skilled in building frameworks with end-to-end automation.

PROFESSIONAL EXPERIENCE:

Organization	Designation	Duration
EXL Services India Pvt. Ltd	Data Scientist (Lead Assistance Manager)	Oct 2021 to till
Nanobi Analytics Pvt.Ltd	Jr. Data Scientist	May 2020 to Oct 2021
Valuestream Business Solution Pvt. Ltd	Data Science Consultant (Contract)	Aug 2019 to Feb 2020

EDUCATION:

University	Course	Duration
Chandigarh University	Master of Science in Data Science	Aug 2024 to Jul 2025
Dr. A.P.J Abdul Kalam University	Bachelor Of Technology in Computer Science and Engineering	Jul 2021 to Jul 2024

KEY EXPERIENCES AND SKILLSETS:

- Good scripting and programming skills in **Python**, **R**, **SAS, MySQL** and **Excel & Advanced Excel**.
- Proficient experience in **Exploratory Data Analysis** (Data Extraction, Transformation, Loading, Data Wrangling, Data Exploration and Data Mining).
- Extensive experience developing different Statistical Machine Learning, Data mining solutions to various business problems and generating data visualizations using **R**, **Python** and **Tableau**.
- Experience in statistical methods **Descriptive Statistic**, **Inferential Statistic**, **Hypothesis Testing**, **Analysis of Variance (ANOVA)** etc.
- Predictive Modelling, Machine Learning, Text Mining & Analytics, Sentiment Analysis, Data Driven Strategy Development through **Visualization Tools**, **Customer Segmentation**, **Risk Analytics**.
- Virtuous knowledge in machine learning techniques and algorithms, such as **Linear Regression**, **Logistic Regression**, **K-NN**, **k-mean Clustering**, **Naive Bayes, SVM** and **Decision tree**, **Random Forest** etc.
- Familiarity with Data Visualization using Matplotlib, Seaborn, Pandas built in Visualization, Plotly, and ggplot2 and **Tableau**.
- Good industry knowledge, analytical and problem-solving skills and ability to work well within a team as well as an individual.
- Able to do **Big Data** Analysis on **AWS PySpark** environment.

PROJECT 1 - Fraud Detection for underwriting modelling

Role: Data Scientist
Organization: Exl Services India

Description: Develop a machine learning model to identify fraudulent applications during the underwriting process, ensuring the integrity of approvals and reducing potential losses. The dataset was prepared with various sources from Transunion and Experian attributes various features considering transactional, behavioral, historical key parameters. Applied various machine learning algorithms (e.g., Logistic Regression, Decision Trees, Random Forest, XGBoost) and selected the best model based on performance metrics. The model provided critical insights for improving the underwriting process and preventing fraud.

Objectives:
Data Analysis & Reporting
Data Modelling
Data Visualization
Predictive Analytics

Responsibilities:
- Modeling- Integrated data from various sources, including TransUnion and Experian, ensuring comprehensive feature engineering using transactional, behavioral, and historical parameters. Performed various data quality checks, missing values treatment, improved the imbalance dataset with SMOTEEN.
- Applied and evaluated multiple machine learning algorithms (e.g., Logistic Regression, Decision Trees, Random Forest, XGBoost) to select the best performing model. Logistic Regression was selected as the final model based on various performance key parameters.
- It improved the fraud detection rate by 21% as compared to exiting models in production.

Environment/Tools: Python (NumPy, Pandas, Sklearn, SciPy, Matplotlib, Seaborn), Trino, Hive, AWS, Pyspark.

PROJECT 2 - Model Monitoring

Role: Data Scientist.
Organization: Exl Services India Pvt Ltd

Description: Developed an automated framework end to end in Python to track and monitor live machine learning models in production for personal loan, purchase finance, auto loan, marketing and loss forecasting models. This framework stores statistical KPI metrics like KS, Gini, ROC, AUC, PSI, CSI & VDI etc. in database. Integrated results with Tableau to create dynamic dashboards for real-time monitoring and reporting. Provided insights to model owners or business stakeholders, recommending recalibration or retraining if model performance deviated from benchmarks consecutive in three monitoring cycle.

Objectives:
- Data Analysis & Reporting
- Dashboard Creation & Monitoring
- Data Visualization
- Predictive Analytics
- KPI Tracking
- Ad-hoc Reporting

Responsibility –
Developed Monitoring Framework- Developed an end-to-end model monitoring framework in Python, SAS, R which tracks statical performance, stability KPI metrics like KS, Gini, ROC, AUC, PSI, CSI & VDI etc. for various models like under writing personal loan, purchase finance, auto loan, marketing model and fraud model in production. Framework developed can be applied to different types of models.
Monitoring – Creating a report tracking the model performance for stack holders, MRM, business team and investor. Depending on the different metrics we flagged the models whose performance degraded for three consecutive monitoring period. We either retard the mode or it goes through recalibration. We track 38 different production models annually where monitoring frequency varies quarterly, half yearly or annually depending on their risk rating of the models.

Environment/Tools: Python (NumPy, Pandas, Sklearn, SciPy, Matplotlib, Seaborn), SAS, R, Trino, Hive, AWS, Pyspark

PROJECT 3 - Fair Lending Project

Role: Data Scientist
Organization: Exl Services India Pvt Ltd

Description – Recalibrate the existing lending model to ensure fairness and compliance with regulations by identifying and removing discriminatory attributes base on race, gender & age and finding alternative attributes that maintain model performance. Conduct thorough data analysis and preprocessing, recalibrated model while using techniques to reduce bias, and evaluate using both standard performance and fairness metrics. Continuously monitor the model's fairness and performance, and provide detailed reports to stakeholders, ensuring an equitable lending process.

Responsibility –
Model Recalibration- We find the attribute causing discrimination/ bias in the existing model using metrics SMD (Standard Mean Deviation). We prepared the dataset from different sources and collected 1458 overall features. Find alternative attributes which keep the balance between performance and bias which require iteratively adding and dropping the attribute identified with equivalent IV values, Clustered attributes. We iteratively ran multiple Logistic Regression with identifier attribute to maintain the actual model performance and reducing the biasness at same time.

Environment/Tools: Python (NumPy, Pandas, Sklearn, SciPy, Matplotlib, Seaborn), Trino, Hive, AWS, Pyspark

PROJECT 4 - Creation and Deployment of a Universal Diagnostic Solution for BI Products

Role: Jr. Data Scientist
Organization: Nanobi Analytics Pvt.Ltd, Bengaluru, India

Description – Developed a Generic diagnostic analytical tool in Python and integrated in a BI product using Django REST API framework. Tools includes feature like prescriptive analysis, descriptive analysis and their interpretation in layman words, Time Series Analysis, Market Basket Analysis, RFM, Clustering, Predictive analytics. Then results are visualized through charting libraries (Fusion Chart, D3 Chart, Tabulator, Bokeh) and charts are rendered through REST API build in Django Rest Framework. Framework was deployed on Windows & Ubuntu server using Nginx, Gunicorn tool.

Responsibility –
Development & Integration- Developed of tools with all features are done in Python Django Rest Frame where we developed the API consuming the connection string and key attributes of the data on which the solution to be built. In backend the logical preprocessing is done, and output is sent over the API to an external system which renders the results using visualization libraries like Fusion Chart, D3 Chart, Tabulator. Framework was deployed on Windows & Ubuntu server using Nginx, Gunicorn tool.

Environment/Tools: Python (NumPy, Pandas, Sklearn, SciPy, Matplotlib, Seaborn), SQL SERVER, Django Rest Framework, Nginx, Gunicorn.

PERSONAL INFORMATION:

Date of Birth: 24th March 1996
Languages known: English, Hindi
Address: D65/269-B, Lahartara, Boulia, Varanasi, Uttar Pradesh, India

Sincerely,

GAURAV GUPTA