Divya Lakshmi

Data Scientist
(940) 252-4873 | divyalakshmi.engr@gmail.com

SUMMARY

- Data Scientist with over 5+ years' experience in Data Extraction, Modelling, Wrangling, Statistical Modeling, and Machine Learning, with recent focus on Generative AI (GenAI) and Large Language Models (LLMs).
- Extensive domain experience in risk analytics, including fraud detection and credit risk modeling, developed in fast-paced fintech environments.
- Domain knowledge and experience in fintech, healthcare, e-commerce, and software industries.
- Expertise in transforming business resources and requirements into manageable data formats and analytical models, designing algorithms, building models, developing data mining and reporting solutions that scale across a massive volume of structured and unstructured data.
- Proficient in managing the entire data science project life cycle from research to deployment, implementing CI/CD principles to ensure strong and reproducible model pipelines.
- Proficient in Machine Learning algorithm and Predictive Modeling including Regression Models, Decision Tree, Random Forests, Sentiment Analysis, Naive Bayes Classifier, SVM, Ensemble Models.
- Proficient in Statistical Methodologies including Hypothetical Testing, ANOVA, Time Series, Principal Component Analysis, Factor Analysis, Cluster Analysis, Discriminant Analysis.
- Knowledge on time series analysis using AR, MA, ARIMA, GARCH and ARCH models.
- Knowledge on Natural Language Processing (NLP) algorithm and Text Mining.
- Worked in large scale database environments like Hadoop and MapReduce, with working mechanisms of Hadoop clusters, nodes and Hadoop Distributed File System (HDFS).
- Strong experience with Python to develop analytic models and solutions, utilizing PyTorch and TensorFlow for deep learning applications.
- Proficient in Python 2.x/3.x with SciPy Stack packages including NumPy, Pandas, SciPy, Matplotlib and IPython .
- Working experience in the Hadoop ecosystem and Apache Spark framework such as HDFS, MapReduce, HiveQL, SparkSQL, PySpark.
- Very good experience and knowledge in provisioning virtual clusters under AWS cloud which includes services like EC2, S3, and EMR.
- Proficient in data visualization tools such as Tableau, Python Matplotlib, R Shiny to create visually powerful and actionable interactive reports and dashboards.
- Excellent Tableau Developer, expertise in building, publishing customized interactive reports and dashboards with customized parameters and user - filters using Tableau(9.x/10.x).
- Experienced in Agile methodology and SCRUM process.
- Strong business sense and abilities to communicate data insights to both technical and nontechnical clients.

WORK EXPERIENCE

Data Scientist **Jun 2024 - Present**
Adobe **Chicago, IL**

- Developed and deployed machine learning models into production using CI/CD pipelines (e.g., Jenkins/GitHub Actions) to automate testing and deployment, ensuring model reliability and performance.
- Collaborated with data engineers to implement ETL process, wrote and optimized SQL queries to perform data extraction from both relational (Amazon Redshift) and NoSQL databases to fit the analytical requirements.

- Explored applications of Generative AI and fine-tuned Large Language Models (LLMs) for internal knowledge retrieval and content summarization tasks.
- Built deep learning models for NLP tasks using PyTorch and TensorFlow, improving model accuracy for sentiment analysis and customer intent classification.
- Utilized Pydantic for data validation and serialization in model inference APIs, ensuring data integrity and reducing runtime errors.
- Performed data analysis by using Hive to retrieve the data from Hadoop cluster, SQL to retrieve data from RedShift.
- Explored and analyzed the customer specific features by using Spark SQL.
- Performed univariate and multivariate analysis on the data to identify any underlying pattern in the data and associations between the variables.
- Performed data imputation using Scikit-learn package in Python.
- Participated in features engineering such as feature intersection generating, feature normalize and label encoding with Scikit-learn preprocessing.
- Used Python 3.X (numpy, scipy, pandas, scikit-learn, seaborn) and Spark 2.0 (PySpark, MLlib) to develop a variety of models and algorithms for analytic purposes.
- Developed and implemented predictive models using machine learning algorithms such as linear regression, classification, multivariate regression, Naive Bayes, Random Forests, K-means clustering, KNN, PCA and regularization for data analysis.
- Conducted analysis on assessing customer consuming behaviors and discovering value of customers with RMF analysis; applied customer segmentation with clustering algorithms such as K-Means Clustering and Hierarchical Clustering.
- Built regression models include: Lasso, Ridge, SVR, XGboost to predict Customer Life Time Value.
- Built classification models include: Logistic Regression, SVM, Decision Tree, Random Forest to predict Customer Churn Rate.
- Used F-Score, AUC/ROC, Confusion Matrix, MAE, RMSE to evaluate different Model performance.
- Designed and implemented recommender systems which utilized Collaborative filtering techniques to recommend courses for different customers and deployed to AWS EMR cluster.
- Utilized natural language processing (NLP) techniques to optimize Customer Satisfaction.
- Designed rich data visualizations to model data into human-readable form with Tableau and Matplotlib.

Data Scientist **Mar 2023 - May 2024**
Stripe **Denton, TX**

- Developed and deployed machine learning models for real-time fraud detection, tackling highly imbalanced datasets using undersampling, oversampling with SMOTE, and cost-sensitive algorithms.
- Built credit risk models to predict potential loan default and loss, implementing ensemble techniques like Ridge, Lasso Regression, and XGBoost.
- Utilized Spark to access and process large-scale data stored in Hadoop (HDFS) and NoSQL data stores like HBase for real-time analysis.
- Conducted deep-dive customer analytics to identify patterns and behaviors associated with fraudulent transactions.
- Wrote complex Spark SQL queries for data analysis to meet business requirements.
- Developed MapReduce/Spark Python modules for predictive analytics & machine learning in Hadoop on AWS; managed codebase and collaboration through GitHub.
- Worked on data cleaning and ensured data quality, consistency, integrity using Pandas, Numpy.
- Participated in feature engineering such as feature intersection generating, feature normalize and label encoding with Scikit-learn preprocessing.

- Improved fraud prediction performance by using random forest and gradient boosting for feature selection with Python Scikit-learn.
- Performed Naïve Bayes, KNN, Logistic Regression, Random forest, SVM and XGboost to identify whether a loan will default or not.
- Implemented Ensemble of Ridge, Lasso Regression and XGboost to predict the potential loan default loss.
- Used various metrics (RMSE, MAE, F-Score, ROC and AUC) to evaluate the performance of each model.
- Used big data tools Spark (Pyspark, SparkSQL, Mllib) to conduct real time analysis of loan default based on AWS.
- Conducted Data blending, Data preparation using Alteryx and SQL for tableau consumption and publishing data sources to Tableau server.
- Created multiple custom SQL queries in Teradata SQL Workbench to prepare the right data sets for Tableau dashboards. Queries involved retrieving data from multiple tables using various join conditions that enabled efficient optimized data extracts for Tableau workbooks.

Data Analyst/Data Scientist **Aug 2020 - Nov 2022**
Ebay **India**

- Gathered, analyzed, documented and translated application requirements into data models and Supports standardization of documentation and the adoption of standards and practices related to data and applications.
- Performed customer analytics and segmentation to understand purchasing behaviors and drive sales strategies.
- Participated in Data Acquisition with the Data Engineer team to extract historical and real-time data by using Sqoop, Pig, Flume, Hive, MapReduce and HDFS.
- Wrote user defined functions (UDFs) in Hive to manipulate strings, dates and other data.
- Performed Data Cleaning, features scaling, features engineering using pandas and numpy packages in Python.
- Applied clustering algorithms i.e. Hierarchical, K-means using Scikit and Scipy.
- Performs complex pattern recognition of automotive time series data and forecast demand through the ARMA and ARIMA models and exponential smoothening for multivariate time series data.
- Delivered and communicated research results, recommendations, opportunities to the managerial and executive teams, and implemented the techniques for priority projects.
- Designed, developed and maintained daily and monthly summary, trending and benchmark reports repository in Tableau Desktop.
- Generated complex calculated fields and parameters, toggled and global filters, dynamic sets, groups, actions, custom color palettes, statistical analysis to meet business requirements.
- Implemented visualizations and views like combo charts, stacked bar charts, pareto charts, donut charts, geographic maps, spark lines, crosstabs etc.
- Published workbooks and extract data sources to Tableau Server, implemented row-level security and scheduled automatic extract refresh.

BI Developer/Data Analyst **Sep 2019 - July 2020**
Optum **India**

- Used SSIS to create ETL packages to Validate, Extract, Transform and Load data into Data Warehouse and Data Mart.
- Maintained and developed complex SQL queries, stored procedures, views, functions and reports that meet customer requirements using Microsoft SQL Server.
- Created Views and Table-valued Functions, Common Table Expression (CTE), joins, complex subqueries to provide the reporting solutions.
- Optimized the performance of queries with modification in T-SQL queries, removed the unnecessary columns and redundant data, normalized tables, established joins and created indexes.

- Created SSIS packages using Pivot Transformation, Fuzzy Lookup, Derived Columns, Condition Split, Aggregate, Execute SQL Task, Data Flow Task and Execute Package Task.
- Migrated data from SAS environment to SQL Server via SQL Integration Services (SSIS).
- Developed and implemented several types of Financial Reports (Income Statement, Profit& Loss Statement, EBIT, ROIC Reports) by using SSRS.
- Developed parameterized dynamic performance Reports (Gross Margin, Revenue based on geographic regions, Profitability based on web sales and smartphone app sales) and ran the reports every month and distributed them to respective departments through mailing server subscriptions and SharePoint server.
- Designed and developed new reports and maintained existing reports using Microsoft SQL Reporting Services (SSRS) and Microsoft Excel to support the firm's strategy and management.
- Created sub-reports, drill down reports, summary reports, parameterized reports, and ad-hoc reports using SSRS.
- Used SAS/SQL to pull data out from databases and aggregate to provide detailed reporting based on the user requirements.
- Used SAS for pre-processing data, SQL queries, data analysis, generating reports, graphics, and statistical analyses.
- Provided statistical research analyses and data modeling support for mortgage products.
- Perform analyses such as regression analysis, logistic regression, discriminant analysis, cluster analysis using SAS programming.

TECHNICAL SKILLS

- **Databases:** MySQL, PostgreSQL, Oracle, NoSQL (MongoDB, DynamoDB, Cassandra, HBase, Redis, Neo4j), Amazon Redshift, MS SQL Server, Taradata
- **Machine Learning/Deep Learning Frameworks:** Scikit-learn, PyTorch, TensorFlow, XGBoost, Spark MLlib
- **MLOps & Tools:** CI/CD Principles, CI/CD Pipelines (Jenkins/GitHub Actions), GitHub, Pydantic, MLflow
- **Big Data Ecosystem:** Hadoop 2.x, Spark 2.x, MapReduce, Hive, Presto, HDFS, Sqoop, Flume
- **Languages:** Python (2.x/3.x), R, SAS, SQL, T-SQL
- **Cloud & Visualization:** AWS (EC2, S3, EMR), Tableau, Matplotlib, Seaborn, ggplot2
- **Statistical Methods:** Hypothetical Testing, ANOVA, Time Series, Confidence Intervals, Bayes Law, Principal Component Analysis (PCA), Dimensionality Reduction, Cross-Validation, Auto-correlation
- **Machine Learning:** Regression analysis, Bayesian Method, Decision Tree, Random Forests, Support Vector Machine, Neural Network, Sentiment Analysis, K-Means Clustering, KNN and Ensemble Method, NLP
- **Operating Systems:** PowerShell, UNIX/UNIX Shell Scripting (via PuTTY client), Linux and Windows

EDUCATION

Master of Science, Computer Science **-** University of North Texas, Denton, TX