Anil Kumar Ojha

Summary

- Data scientist with 5+ years of experience in transforming business requirements into actionable data models, prediction models and informative reporting solutions.
- Employing various SDLC methodologies such as Agile and SCRUM methodologies.
- Expert in the entire Data Science process life cycle including Data Acquisition, Data Preparation, Data Manipulation, Feature Engineering, Machine Learning Algorithms, Validation and Visualization.
- Hands - on experience in Machine Learning algorithms such as Linear Regression, GLM, CART, SVM, KNN, LDA/QDA, Naive Bayes, Random Forest, SVM, Boosting,
- Proficient in Python and its libraries such as NumPy, Pandas, Scikit-learn, Matplotlib and Seaborn.
- Experience in working with Hadoop Big Data tools such as HDFS, Hive, Pig Latin and Spark.
- Expertise in implementing writing and optimizing the HiveQL queries.
- Deep knowledge of SQL languages for writing Queries, Stored Procedures, User-Defined Functions, Views, Triggers, Indexes and etc.
- Designed, trained, validated end to end ML models & pipelines for image, video, timeseries, seq2seq using TensorFlow, Keras, Pytorch to setup deep CNN.
- Maintenance and monitoring of Docker in a cloud-based service during production and Set up a system for dynamically adding and removing web services from a server using Docker.
- Worked on integration of diverse mathematical and statistical procedures, pattern recognition, model building, creating various scientific and industrial packages within R.
- Knowledge and experience in agile environments such as Scrum and using project management tools like Jira/Confluence and version control tools such as GitHub/Git.
- Collaborated with data engineers to implement ETL process, wrote and optimized SQL queries to perform data extraction from Cloud and merging from Oracle.
- Hands-on experience in importing and exporting data using Relational Database including Oracle, MySQL and MS SQL Server, and NoSQL database like MongoDB.
- Experience in implementing data analysis with various analytic tools, such as Jupyter, Notebook, Spyder, Reshape, ggplot2, Dlpr, Car, Mass, SAS,

Matlab and Excel.

- Good team player and quick-learner; highly self-motivated person with good communication and interpersonal skills.

Skills

Languages	Python, R, SQL
Databases	SQL Server, MS-Access, Oracle 11g/10g/9i and Teradata, big data, Hadoop
DWH / BI Tools	Microsoft Power BI, Tableau, SSIS, SSRS, SSAS, Business Intelligence Development Studio (BIDS), Visual Studio, Crystal Reports.
Big Data technologies	Hadoop, Hive, HDFS, MapReduce, Pig, Kafka.
Tools and Utilities	SQL Server Management Studio, SQL Server Enterprise Manager, SQL Server Profiler, Import &Export Wizard, Visual Studio .Net, Microsoft Management Console, Visual Source Safe 6.0, DTS, Crystal Reports, Power Pivot, ProClarity, Microsoft Office, Excel Power Pivot, Excel Data Explorer
Machine Learning	Linear Regression, Logistic Regression, Gradient boosting, Random Forests, Maximum likelihood estimation, Clustering, Classification & Association Rules, K-Nearest Neighbors (KNN), K-Means Clustering, Decision Tree (CART &CHAID), Neural Networks, Principal Component Analysis, Weight of Evidence (WOE) and Information Value (IV), Factor Analysis, Sampling Design, Time Series Analysis, ARIMA, ARMA, GARCH, Market Basket Analysis, Text mining
Cloud Technologies:	Amazon Web Services (EC2, EBS, S3, VPC, RDS, S3, SES)
Methodologies	Agile, Scrum

Experience

Synchrony Financial, Stamford, CT **September**

2018 to Present
Role: Data Scientist

Description: Synchrony Financial is a consumer financial services company headquartered in Stamford, Connecticut, United States. The company offers consumer financing products, including credit, promotional financing and loyalty programs, instalment lending, and FDIC insured savings products through Synchrony Bank, its wholly owned subsidiary.

Responsibilities:

- Built data pipelines for reporting, alerting, and data mining, table design and data management using HDFS, Hive, Impala, Sqoop, MySQL, Mem SQL, Grafana/Influx DB, and Kafka.
- Built models using Statistical techniques like Bayesian HMM and Machine Learning classification models like XG Boost, SVM, and Random Forest using R and Python packages.
- Worked with data compliance teams, data governance team to maintain data models, Metadata, data Dictionaries, define source fields and its definitions.
- Working in Business and Data Analysis, Data Profiling, Data Migration, Data Integration and Metadata Management Services.
- Worked extensively on Databases preferably Oracle 11g/12c and writing PL/SQL scripts for multiple purposes. Worked with Big Data Technologies such Hadoop, Hive, MapReduce.
- Developed MapReduce/Spark Python modules for machine learning & predictive analytics in Hadoop on AWS. Implemented a Python-based distributed random forest via Python streaming.
- Worked with statistical models for data analysis, predictive modelling, machine learning approaches, and recommendation and optimization algorithms.
- A highly immersive Data Science program involving Data Manipulation & Visualization, Web Scraping, Machine Learning, Python programming, SQL, GIT, Unix Commands, NoSQL, MongoDB, Hadoop.
- Data analysis using regressions, data cleaning, excel v-look up, histograms and TOAD client and data representation of the analysis and suggested solutions for investors
- Attained good knowledge in Hadoop Data Lake Implementation and HADOOP Architecture for client business data management.
- Performed scoring and financial forecasting for collection priorities using Python, R and SAS machine learning algorithms. Developed a legal model for predicting which debtors respond to litigation only.
- Handled importing data from various data sources, performed transformations using Hive, MapReduce, and loaded data into HDFS
- Managed existing team members, lead the recruiting and on boarding of a larger Data Science team that addresses analytical knowledge requirements.
- Created SQL scripts and analysed the data in MS Access/Excel and Worked on SQL and SAS script mapping.
- Worked directly with upper executives to define requirements of scoring models.
- Developed a model for predicting a debtor setting up a repayment rehabilitation program for student loan debt. Setup storage and data analysis

tools in Amazon Web Services cloud computing infrastructure.
- Rapid model creation in Python using pandas, numpy, sklearn, and plot.ly for data visualization. These models are then implemented in SAS where they are interfaced with MSSQL databases and scheduled to update on a timely basis.
- Above scoring models resulted in millions of dollars of added revenue to the company and a change in priorities of the entire company.

Environment: R, SQL, Python 2.7.x, SQL Server 2014, regression, logistic regression, random forest, neural networks, Topic Modelling, NLTK, SVM (Support Vector Machine), JSON, XML, HIVE, HADOOP, PIG, Sklearn, SciPy, Graph Lab, No SQL, SAS, SPSS, Spark, Hadoop, Kafka, HBase, MLib.

Bonobos, Inc - New York, NY **August 2016 to September 2018**
Data Scientist

Description: Bonobos is an e-commerce apparel company that designs and sells men's clothing and it's a data driven company doing business by using advanced machine learning techniques.
The goal of the project was to support sales and marketing team by building machine learning models and providing data analysis on sales forecasting which allows businesses to plan for the future and be prepared to meet demands and maximize profits, help formulate better marketing strategy and worked closely with business team to onboard new model.

Responsibilities:

- Collaborated with data engineers and operation team to implement ETL process, wrote and optimized SQL queries to perform data extraction to fit the analytical requirements.
- Worked on data cleaning and ensured data quality, consistency, integrity using Pandas, NumPy.
- Explored and analyzed the customer specific features by using Matplotlib, Seaborn in Python and dashboards in Tableau.
- Performed data imputation using Scikit-learn package in Python.
- Participated in features engineering such as feature generating, PCA, feature normalization and label encoding with Scikit-learn preprocessing.
- Used Python 2.x/3.X (NumPy, SciPy, Pandas, Scikit-learn, seaborn to develop variety of models and algorithms for analytic purposes.
- Experimented and built predictive models including ensemble methods such as Gradient boosting trees and Neural Network by Keras to predict Sales amount.
- Conducted analysis and patterns on customers' shopping habits in different

location, different categories and different months by using time series modeling techniques.

- Used RMSE/MSE to evaluate different models' performance.
- Designed rich data visualizations to model data into human-readable form with Tableau and Matplotlib.

Environment: Python, SQL, Oracle 12c , PL/SQL, T-SQL, regression, Cluster analysis, Scala NLP, Spark, Kafka, MongoDB, logistic regression, random forest, OLAP, MariaDB, SAP CRM, HDFS, ODS, NLTK, SVM, JSON, Tableau, XML,.

UnitedHealth Group, NJ **June 2014 to August 2016**
Role: Data Scientist

Responsibilities:

- Participated in all phases of data mining, data collection, data cleaning, developing models, validation, and visualization to deliver data science solutions.
- Collected data using Hadoop tools to retrieve the data required for building models such as Hive and Pig.
- Developed Spark Python modules for machine learning & predictive analytics in Hadoop.
- Implemented a Python-based distributed random forest via PySpark.
- Used Pandas, NumPy, Seaborn, Matplotlib, Scikit-learn in Python for developing various machine learning models and utilized algorithms such as Linear regression, Logistic regression, Gradient Boosting, SVM and KNN.
- Designed and developed machine learning models to improve advertising agencies programmatic strategies for optimal biddings of impression opportunities.
- Analyzed and visualized different segments of users to understand their advertisement behaviors better with Tableau.
- Generated comprehensive analytical reports by running SQL queries against current databases to conduct data analysis.
- Work on data pre-processing and cleaning the data to perform feature engineering and performed data imputation techniques for the missing values in the dataset using Python.
- Generated graphs and reports using ggplot package in RStudio for analytical models.
- Used Pandas API for analyzing time series.
- Developed MapReduce pipeline for feature extraction using Hive and Pig.
- Used MS Excel, MS Access and SQL to write and run various queries.

Environment: Hadoop, AWS, GIT, Unix, Python 3.5.2, Spark MlLib, SAS, regression, logistic regression, NoSQL, OLTP, Random Forests, OLAP, HDFS, NLTK, SVM, JSON, XML, MapReduce.

Education: Available upon Request

References: Available upon Request