**Dipendra Karki**

**Summary:**

* Around 5+ years of experience in Data Acquisition, Data Pre-Processing, Data Visualization, Feature Engineering and building unsupervised and supervised Machine Learning models including complex models such as deep neural networks and convolutional neural networks.
* Proficient in building the models for both batch data and stream data as well as fine-tuning the models for robust performance and accuracy.
* Hands on experience in implementing (I) supervised learning (linear and logistic regression, boosted decision trees, Support Vector Machines, neural networks, NLP), (II) unsupervised learning (clustering, dimensionality reduction, recommender systems), (III) probability & statistics, experiment analysis, confidence intervals, A/B testing
* Experience working with Big Data technologies like Apache Hadoop, Spark, Hive, HBase and MongoDB. Hands-on experience with Hadoop ecosystems and HDFS.
* Experience working with cloud platforms like Amazon Web Service (AWS) and Google Cloud Platform (GCP) for building and deploying big data and machine learning solutions to the cloud.
* Good at data structures and algorithm and design techniques. Hands on experience in data mining and machine learning algorithms and approaches.
* Proficient in Python and Java; data visualization, ML and deep learning libraries like Pandas, Numpy, Scikit-learn, Tensorflow, Keras; natural language processing libraries like NLTK and spaCy.
* Hands on experience in implementing PCA, LDA, Naive Bayes and skilled in Decision Trees, Random Forests, Linear and Logistic Regression, SVM, Clustering, neural networks and sound knowledge on Recommender Systems.
* Hands-on experience with integrating and management of data from different sources like Oracle, SQLServer, MySQL and NoSQL databases like MongoDB.
* Experience in designing visualizations using Tableau software on web and desktop platforms, publishing and presenting dashboards.
* Experience in developing Custom Report and different types of Tabular Reports, Matrix Reports, Ad hoc reports and distributed reports in multiple formats using SQL Server Reporting Services (SSRS).
* Experience with Advanced Analytics using MicroStrategy to provide business intelligence solutions through descriptive analytics to facilitate with business actions and plans.
* Experience working with APIs like Twilio for integrating messaging systems with applications.
* Experience with product deployment using containerized systems like dockers.
* Experience with working in a team and leading projects to deliver exceptional results within the desired timeframe, quick learner and always open to learn and adapt new technologies.
* Experience working with Agile and Waterfall models, dealing with sprints and resolving issues within each story of a sprint. More than 3 years of experience working with Agile methodology frameworks like JIRA.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Programming Languages** | Python, Java, C/C++, Scala, JavaScript, SQL |
| **Packages** | **Data Acquisition & Processing**: pandas, numpy, sciPy, tweepy  **Web Scraping**: BeautifulSoup, Selenium, Scrapy  **Data Visualization**: matplotlib, seaborn, ggplot2  **Machine Learning**: scikit-learn, tensorflow, Keras  **NLP**: nltk, spaCy, Stanford NLP |
| **Big Data Technologies** | **Batch Data**: Hadoop, MapReduce, HDFS, Apache Spark  **Streaming Data**: Apache Spark (with Scala and Python-PySpark), Apache Kafka |
| **Machine Learning** | **Supervised**: Linear Regression, Logistic Regression, Support Vector Machine, Decision Trees, Random Forest, Perceptron, Artificial Neural Networks, Back Propagation Algorithms, Recurrent Neural Networks, Convolutional Neural Networks, Deep Learning with Keras and TensorFlow, Ensemble Methods like Boosting, Bagging, Gradient Boosting  **Unsupervised**: k-means clustering  Model Evaluation, Cross-Validation  Hands-On experience with Google Colab |
| **Natural Language Processing** | NLTK toolkit, Regular Expressions, Text Analytics, Text Mining, Sequence Tagging, POS Tagging, Named Entity Recognition, Relation Extraction, Universal Dependency (UD) Parsing (Cross-Linguistic Platform) |
| **BI Tools** | Tableau, Tableau server, Tableau Reader, SAP Business Objects, OBIEE, QlikView, SAP Business Intelligence, Amazon Redshift, or Azure Data Warehouse |
| **Web** | **Frontend**: AngularJS, HTML5, CSS3, Bootstrap  **Backend**: NodeJS, PHP  **Frameworks**: Express, Django, Flask |
| **Databases** | **Relational**: MSSQL, MySQL, PostgreSQL, Oracle, Teradata  **NoSQL**: MongoDB, HBase, Hive, Cassandra  **ORM**: SQLAlchemy |
| **Reporting Tools** | Tableau, SSRS |
| **ETL Tools** | Oracle Data Integrator (ODI), Informatica Power Centre, SSIS |
| **Version Control Tools** | GitHub, GitLab, SVM |
| **Project Execution**  **Methodologies** | Ralph Kimball and Bill Inmon data warehousing methodology, Rational Unified Process (RUP), Rapid Application Development (RAD), Joint Application Development (JAD). |
| **Operating Systems** | Windows, Linux, Unix, Mac OS |

**PROFESSIONAL EXPERIENCE:**

**Role-** Machine Learning Engineer/Data Scientist

**Client-** JC Penny, Plano, Texas

**Project Duration-** Aug 2018 - Present

**Description:**

JC Penny is an American department store chain retail store. The project was to build predictive models for forecasting the demand of products based on the historical sales data available on multiple products.

**Responsibilities**:

* Gathering business requirement from client and approach formulation and design methodology to match client requirements.
* Data extraction by developing a pipeline using AWS data pipelines to retrieve the data from S3 buckets. Also, analyzing the data using Amazon Redshift Spectrum to query terabytes of data present in S3.
* Performed intensive data pre-processing, features scaling, features engineering using Python
* Replacement of missing data and perform a proper Exploratory Data Analysis (EDA) to understand the time series data.
* Checked different components of the time series data like Autocorrelation, Seasonality and Stationarity.
* Designed rich data visualizations using Tableau and matplotlib along with seaborn to present the findings of the analysis to business leaders and executives.
* Involved working in Data science using Python 2.x/3.x on different data transformation and validation techniques like Dimensionality reduction using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), Factor Analysis, testing and validation using ROC plot, K- fold cross validation, statistical significance testing.
* Used Python to develop and train ML models such as Autoregressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), Neural Networks with combination of clustering algorithms using Keras, TensorFlow and Sklearn that helped in forecasting the future sales. CUDA capable GPU was used for intensive parallel processing of huge data.
* Fine tuning and improvise the model until reasonable accuracy is obtained.
* Used Agile methodology and Scrum during project development.

**Environment**: Python 2.x/3.x, AWS(S3/Redshift), AWS data pipelines, TensorFlow, Keras, Scikit-learn, Tableau, Pandas, XML

**Role-** Data Scientist/Machine Learning

**Client-** Charter Communications, Stamford, CT

**Project Duration-** Aug 2016 – Jun 2018

**Description:**

The objective of the project was to classify emails that can be routed to specific departments of the company according to the level of severity thereby eliminating manual work of classification. Used NLTK and spaCy Python libraries for this project in the AWS cloud environment. Worked till Deployment of the model to create APIs that can serve Web requests.

**Responsibilities:**

* A highly immersive Data Science project involving Data Manipulation & Visualization, Machine Learning, Python programming, SQL, GIT, Unix Commands, NoSQL, MongoDB, Hadoop.
* Installed and used TensorFlow Deep Learning Framework
* Worked on different data formats such as CSV, JSON, XML and performed Machine Learning algorithms in Python and Deep Learning techniques such as RNN and LSTM
* Setup storage and data analysis tools in Amazon Web Services cloud computing infrastructure.
* Used pandas, NumPy, seaborn, SciPy, matplotlib, scikit-learn, NLTK in Python for developing various machine learning algorithms.
* Used techniques in NLP like Noise Removal, Lemmatization, Stemming, POS Tagging, Bag of Words, Topic Modelling, TF-IDF, word2vec.
* Development and Deployment using Flask.
* Worked as Data Architect and IT Architect to understand the movement of data and its storage.
* Performed Data Cleaning, features scaling, features engineering using Pandas and NumPy packages in python and build models using deep learning frameworks.
* Implemented application of various machine learning algorithms and statistical modeling like Text Analytics, Sentiment Analysis, Decision Tree, Naive Bayes, Logistic Regression and Linear Regression using Python to determine the accuracy rate of each model.
* Implemented Agile Methodology for building an internal application.
* Extracting the source data from Oracle tables, MS SQL Server, sequential files and excel sheets.
* Developing and maintaining Data Dictionary to create metadata reports for technical and business purpose.
* Developed MapReduce/Spark Python modules for machine learning & predictive analytics in Hadoop on AWS.
* 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.
* Extracted data from HDFS and prepared data for exploratory analysis for data munging.

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

**Role-** Data Scientist/Machine Learning

**Client-** Express, St Louis. MO

**Project Duration-** Jun 2014- Jul 2016

**Responsibilities:**

* Worked with several R packages including knitr, dplyr, SparkR, CausalInfer, Space-Time.
* Coded R functions to interface with Caffe Deep Learning Framework.
* Used Pandas, Numpy, Seaborn, SciPy, Matplotlib, Sci-kit-learn, and NLTK in Python for developing various machine learning algorithms.
* Installed and used CaffeDeep Learning Framework
* Worked on different data formats such as JSON, XML and performed machine learning algorithms in Python.
* Setup storage and data analysis tools in Amazon Web Services (AWS) cloud computing infrastructure.
* Implemented end-to-end systems for Data Analytics, Data Automation and integrated with custom visualization tools using R, Mahout, Hadoop and Mongo DB.
* Worked as Data Architects and IT Architects to understand the movement of data and its storage and ER Studio 9.7.
* Utilized Spark, Scala, Hadoop, HBase, Cassandra, MongoDB, Kafka, Spark Streaming, MLLib, Python, a broad variety of machine learning methods including classifications, regressions, dimensionally reduction etc. and Utilized the engine to increase user lifetime by 45% and triple user conversations for target categories.
* Used Spark Data frames, Spark-SQL, Spark MLLib extensively and developing and designing POC's using Scala, Spark SQL and MLLib libraries.
* Used Data Quality Validation techniques to validate Critical Data Elements (CDE) and identified various anomalies.
* Extensively worked on Data Modeling tools Erwin Data Modeler to design the Data Models.
* Developed various Qlik-View Data Models by extracting and using the data from various sources files, DB2, Excel, Flat Files and Big data.
* Participated in all phases of Datamining, Data-collection, Data-Cleaning, Developing-Models, Validation, Visualization and Performed Gap Analysis.
* Data Manipulation and Aggregation from different source using Nexus, Toad, Business Objects, PowerBI and SmartView.
* Implemented Agile Methodology for building an internal application.
* Focus on integration overlap and Informatica newer commitment to MDM with the acquisition of Identity Systems.
* Good knowledge of Hadoop Architecture and various components such as HDFS, JobTracker, Task Tracker, Name Node, Data Node, Secondary NameNode, and MapReduce concepts.
* As Architect delivered various complex OLAP Databases/Cubes, Scorecards, Dashboards and Reports.
* Programmed a utility in Python that used multiple packages (SciPy, Numpy, Pandas)
* Implemented Classification using supervised algorithms like Logistic Regression, Decision trees, KNN, Naive Bayes.
* Designed both 3NF data models for ODS, OLTP systems and Dimensional Data Models using Star and Snowflake Schemas.
* Updated Python scripts to match training data with our database stored in AWS Cloud Search, so that we would be able to assign each document a response label for further classification.
* Created SQL tables with referential integrity and developed queries using SQL, SQL PLUS and PL/SQL.
* Designed and developed Use Case, Activity Diagrams, Sequence Diagrams, OOD (Object oriented Design) using UML and Visio.
* Interaction with Business Analyst, SMEs and other Data Architects to understand Business needs and functionality for various project solutions
* Interaction with Business Analyst, SMEs, and other Data Architects to understand Business needs and functionality for various project solutions
* Identifying and executing process improvements, hands-on in various technologies such as Oracle, Informatica, and BusinessObjects.

**Environment:** AWS, R, Informatica, Python, HDFS, ODS, OLTP, Oracle 10g, Hive, OLAP, DB2, Metadata, MS Excel, Mainframes MS Vision, Map-Reduce, Rational Rose, SQL, and MongoDB.