**Aashish Khadka**

**Professional Experience:**

* Over 6 years of experience of **Machine Learning**. Excellent capability in collaboration, quick learning and adaptation
* Experience in **Data mining** with large datasets of Structured and Unstructured data, Data Acquisition, Data Validation, Predictive modeling, Data Visualization.
* Experience in integrating data, profiling, validating and data cleansing transformation and data visualization using **R and Python**
* Theoretical foundations and practical hands-on projects related to **(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, (iv) algorithms and data structures**
* Hands on experience in design, management and visualization of databases using **Oracle, MySQL and SQL Server**
* Experience in dimensionality reduction using techniques like **PCA and LDA**
* Intensive hands-on Boot camp on **Data Analytics** course spanning from Statistics to Programming including data engineering, data visualization, machine learning and programming in **R, SQL**
* Good Exposure with **Factor Analysis, Bagging and Boosting algorithms**
* Experience in Descriptive Analysis Problems like Frequent **Pattern Mining, Clustering, Outlier Detection**
* Worked on Machine Learning algorithms like Classification and Regression with **KNN Model, Decision Tree Model, Naïve Bayes Model, Logistic Regression, SVM Model and Latent Factor Model**
* Hands-on experience on **Python** and libraries like **NumPy, Pandas, Matplotlib, Seaborn, NLTK, Sci-Kit learn, SciPy**
* Good Knowledge on **Natural Language Processing (NLP)**, **Anomaly Detection**, **Time Series Analysis** and Forecasting using **ARIMA model** in **Python and R**
* Good knowledge in **Tableau, Power BI** for interactive data visualizations
* 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)**

**Technical set:**

|  |  |
| --- | --- |
| **Programming Languages** | Python (2 &3), R, SAS, SQL, SQL Script |
| **Machine Learning** | Regression analysis, Decision Tree, Random Forests, Naïve Bayes, SVM, Neural Network, Sentiment Analysis, K-Means Clustering and KNN, NLP, Time Series Analysis |
| **Frameworks and Toolkits** | NumPy, Pandas, Matplotlib, Scikit-learn, SciPy, Seaborn, Plotly, StatsModel |
| **Statistical Methods** | Descriptive statistics, Hypothesis Testing, ANOVA, Chi-square testing, F-tests, Confidence Intervals, Bayes Law, Dimensionality Reduction, ROC curve |
| **Databases** | Oracle 10g/9i/8i, MS SQL Server, MS Access, MySQL 5.x, Teradata, MongoDB, Hadoop |
| **Operating Systems** | Windows NT/2000/XP/7,8&10, Windows Server 2008 |

|  |  |
| --- | --- |
| **Client: BBVA Compass, Dallas, Texas** | **Jul 2017 – Feb 2019** |
| **Role: Data Scientist / Machine Learning Engineer** |  |

**Project Description:** Building a robust Anti-Money Laundering Compliance Technology using Big Data Analytics, Machine Learning and AI. Project encompasses five major areas of AML Compliance i.e. Automating collection of Due Diligence data of KYC using APIs, Using Text Analytics to process documents in Trade Finance, Implementing Advanced Entity matching to create Single(360) View of a Customer, Bringing transactions data, case management data, customers data and external data in one place and using Machine Learning Algorithms to uncover hidden patterns. Objectives of the project are: Detect AML violations on a proactive basis; Save on staffing expenses for Due Diligence; Increase accurate production of suspicious activity reports (SAR); Help create Customer 360 that can help accelerate Customer Segmentation.

**Responsibilities:**

* Extracted data from **HDFS** and prepared data for exploratory analysis using **data munging**
* Built models using Statistical techniques like Bayesian HMM and Machine Learning classification models like **XG Boost**, **SVM**, and **Random Forest**
* Participated in all phases of **data mining, data cleaning, data collection, developing models, validation, visualization,** and **performed Gap analysis**
* Extracted data from **HDFS** and processed obtained data using **Pig**, **Hive** and **Apache Spark**
* A highly immersive Data Science program involving **Data Manipulation** &**Visualization**, **Web Scraping**, **Machine Learning, Python programming, SQL**, **Mongo DB, Hadoop**
* Setup storage and data analysis tools in **AWS cloud computing infrastructure**
* Worked on different data formats such as **JSON, XML** and performed machine learning algorithms in **Python**
* Used **pandas, numpy, sea born, matplotlib, scikit-learn, scipy, NLTK** in **Python**
* Programmed a utility in Python that used multiple packages (**numpy, scipy, pandas**)
* Worked on building an **Anomaly Detection** algorithm to monitor transactions in real time
* Implemented PCA Dimensionality reduction and un**s**upervised classification algorithms like

**K-Means Clustering, Hierarchical Clustering, Probabilistic Clustering**

* Data transformation from various resources, data organization, features extraction from raw and stored
* Validated the machine learning classifiers using **ROC Curves** and **Lift Charts**
* Validated clustering algorithms using **pair-wise precision** and **recall**, **f-score**, **SSE criterion**

|  |  |
| --- | --- |
| **Client: USAA, San Antonio, TX** | **April 2015 – May 2017** |
| **Role: Data Scientist** |  |

**Project Description:** Building a scalable loan-default risk prediction model using Machine Learning Algorithms. Models developed should be robust in dealing with imbalanced data related to loan default. Objective is to improve credit scoring and help bank lower loss by lowering possible default status as well as lowering rate of false positives.

**Responsibilities:**

* Perform ad-hoc exploratory statistics and data mining tasks on small scale to big data
* Delivered an interactive dashboard in Tableau to visualize 8 billion rows (1.2 TB) loan data
* Performed Data Analysis and Data Profiling and worked on data transformations and data quality rules
* Built credit risk scorecards and marketing response models using SQL and SAS.
* Created predictive models to analyze the behavior of customer in purchase using Python and R
* Performed data Cleaning, features scaling, features selection using Pandas, NumPy and scikit-learn packages in python
* Used Principal Component Analysis in feature engineering to analyze high dimensional data
* Applied Random Forest Classification, KNN, SVM, Logistic regression and Ensemble Machine Learning models to classify customer as risky and non-risky customers
* Using graphical packages produced ROC Curve to visually represent True Positive Rate versus False Positive Rate. Equally produced visualization of Precision Recall Curve for Area under the Curve
* Objective achieved was 84% accuracy in predicting default on out-of-sample using Logistic Regression

|  |  |
| --- | --- |
| **Client: AT&T- Houston, TX** | **Feb 2013-April 2015** |
| **Role: Data Analyst** |  |

**Project Description:** Support Data Science team in building best model for customer churn prediction. Building dynamic data pipeline and act as a bridge between Data Engineers and Data Scientists to make sure that appropriate data is used for modeling.

**Responsibilities:**

* As part of this team I was involved in Retrieve data from Database and developing reports based on the business requirements
* Generating the reports for Teradata and Oracle Database to analyze the customer behavior and plan strategies to improve the response rates of marketing Campaigns
* Improved performance while using performance tuning techniques like primary Index and collection statistics on index columns
* Strong Experience in writing Python programs for manipulating the data reading from various Teradata and convert them as one CSV Files
* Strong experience in using Excel & MS Access to dump the data & analyze based on business needs
* Performed Verification, Validation and Transformations on the Input data (Text files) before loading into target database
* Responsible for analyzing business requirements and developing Reports using PowerPoint, Excel to provide data analysis solutions to business clients
* Developed and automated the data manipulation process for above using stored procedures/views in SQL Server
* Integrated various relational and non-relational sources such as DB2, Oracle, SQL Server, XML and Flat Files
* Performed administrative tasks, including creation of database objects such as database, tables, and views, using SQL DCL, DDL, and DML requests
* Created Data Quality Scripts using SQL to validate successful data load and quality of the data