Prayash Baral

**Professional Summary:**

* **6** years of experience in **Data Science, Machine Learning, Data Mining, Data Acquisition, Data Validation, Predictive Modelling** and **Data Visualization.**
* Expertise in transforming business requirements into **analytical models, designing algorithms, building models**, developing **data mining** and reporting solutions that scales across massive volume of **structured and unstructured data.**
* Extensive experience in **Text Analytics**, developed a text processing module to help with entity matching**, text categorization/routing** and **named-entity extraction.**
* Hands on experience with developing different **Statistical Machine Learning, Data Mining** solutions to business problems and generating **data visualizations** using **Python** and **Tableau**.
* Experience with **SparkMlib** utilities to perform task such as **Classification**, **Regression**, **Clustering**, **Collaborative Filtering, Dimensionality Reduction**.
* Extensive experience in building data models by using Machine learning techniques such as **Cluster Analysis**, **Market Based Analysis**, **Association Rules** and **Naive Bayes technique**.
* Hands on experience in implementing **Naïve Bayes, Random Forests, Decision Trees, Linear** and **Logistic Regression, SVM, Clustering, Neural Networks, Principle Component Analysis (PCA)** and good knowledge on **Recommender Systems.**
* Adept in programming language like **Python** including **Big Data** technologies like **Spark, Hive**.
* Skilled in using **pandas** in python for performing exploratory **data analysis**.
* Experience in designing stunning **visualizations** using **Tableau** software and publishing and presenting **dashboards**, **Storyline** on web and desktop platforms.
* Experience in developing **ad-hoc reports**, scheduling the processes and administering the **tableau** activities.
* Experienced in designing customized interactive dashboards in **Tableau** using **Marks**, **Action**, **Filters**, **Parameter** and **Calculations**.
* Extensive experience with **advanced mathematics, statistics**, and **visualizing complex data** and concepts to diverse audiences.
* Skilled in using **Hadoop** (**Pig** and **Hive**) for basic **analysis** and extraction of **data** in the infrastructure to provide **data summarization**.
* Ability to work in **tight schedules** and efficient in meeting **deadlines**.
* Excellent **Analytical, communication** and **interpersonal** skills.
* Experience in developing application using **Agile Scrum methodology**.
* **Adapt** and **adhere** to **industry** standard while working with **multi**-**cultural** **teams**.
* Hands on experience in working with **UML**, **Class Diagram** and **Sequence diagram**.
* Experience in **implementing application** using different IDEs like **PyCharm, Spyder** and **Jupyter** **Notebook**.

**TOOLS AND TECHNOLOGIES:**

|  |  |
| --- | --- |
| **Machine Learning** | **Regression (Linear, Logistic ) analysis, Decision Tree, Random Forests, K-Means, Naïve Bayes, SVM, Neural Network(CNN, RNN), K-Means Clustering and KNN.** |
| **Programming Languages** | **Python , Java, C/C++, SQL** |
| **Frameworks** | **Django, Flask** |
| **Libraries** | **NumPy, Pandas, NLTK, Scikit-learn,** |
| **Statistical Methods** | **Descriptive statistics, Hypothesis Testing, ANOVA, Chi-square testing,**  **F-tests, Confidence Intervals, Bayes Law, Dimensionality Reduction,**  **ROC curve.** |
| **Databases** | **Oracle 10g/9i, MS SQL Server 2008, MS Access, MySQL 5.7, PostgreSQL, MongoDB, Cassandra** |
| **Tools & IDEs** | **Tableau, Virtual Box, GitLab, Sublime Text, Notepad++, PyCharm, Sypder, Jupyter Notebook, Slack, Microsoft Team** |
| **Operating Systems** | **Windows 7,8&10, Windows Server 2008, Linux** |

|  |
| --- |
| **Client: Sprint, Kansas City, KS** |
| **Role: Data Scientist/ Machine Learning Engineer June 2017 - Mar 2019** |
|  |

**Description:**

Sprint Corporation is an American [telecommunications](https://en.wikipedia.org/wiki/Telecommunications) [company](https://en.wikipedia.org/wiki/Company) that provides [wireless services](https://en.wikipedia.org/wiki/Wireless_telecommunications) and is an [internet service provider](https://en.wikipedia.org/wiki/Internet_service_provider). The purpose of this project is to develop predictive models to measure success of business events based on user involvement and feedbacks and present the results to business partners and customer relationship management group.

**Responsibilities:**

* Responsible for Data Cleaning, features scaling, features engineering by using NumPy and Pandas in Python.
* Performed Exploratory Data analysis (EDA) to maximize insight in to the dataset, detect the outliners and extract important variables by graphically and Numerically.
* Developed various Clustering algorithms for market segmentation to analyze the customer behavior patterns.
* Used Pandas, NumPy, seaborn, SciPy, Matplotlib, Seaborn, Scikit-learn, NLTK in Python at various stages for developing machine learning model.
* Implemented machine learning algorithms, Random forest and Support vector machines to predict the Customer churn and Customer interface.
* Used cross-validation techniques to avoid the overfitting of the model to make sure the predictions are accurate and measured the performance using Confusion matrix and Classification report.
* Improved accuracy using Ensemble methods of the training model with different Bagging and Boosting methods.
* Performed Text analytics on unstructured email data using Natural language processing tool kit(NLTK).
* Involved in various pre-processing phases of text data like Tokenizing, Stemming, Lemmatization and converting the raw text data to structured data
* Performed feature engineering, performed NLP by using some techniques like Word2Vec, BOW (Bag of Words), tf-idf, Avg-Word2Vec, if-idf, Weighted Word2Vec.
* Used PySpark Machine learning library to build and evaluate different models.
* Used Tableau to convey the results by using dashboards to communicate with team members and with other data science teams, marketing and engineering teams.
* Communicated the results with operations team for taking best decisions.

**Environment:** NumPy, Pandas, Matplotlib, Seaborn, Scikit-Learn, Tableau, SQL, Git, Microsoft Excel, PySpark-ML, Random Forests, SVM, PCA, K-Means, Natural Language Tool Kit.

|  |
| --- |
| **Client: Vertiv, Columbus, OH** |
| **Role: Data Scientist March 2015 - May 2017** |
|  |

**Description:** Vertiv is an Ohio-based provider of equipment and services for datacenters. The purpose of the project in Vertiv was to identify and develop advanced analytical models, statistical models, machine learning methods and solutions for operation clients to improve various business outcome indicators.

**Responsibilities:**

* Implemented end-to-end systems for Data analytics, Data Automation and customized visualization tools using Python, Hadoop and MongoDB.
* Used Pandas, NumPy, Seaborn, Matplotlib, scikit-learn in Python for developing various machine learning algorithms.
* Worked on csv, Json, excel different types of files for the Data cleaning and Data analysis
* Used Python for statistical operations on the data and Matplotlib for the visualizing the data
* Ensured that the model has a low False Positive Rate.
* Managed large datasets using Pandas data frames and MySQL.
* Built various graphs for business decision-making using Python Matplotlib library.
* Identified root causes of problems, and facilitated the implementation of cost-effective solutions with all levels of management.
* Performed Data Cleaning, handled missing data, outliers, feature scaling, features engineering.
* Application of various ML algorithms and statistical modeling like decision trees, regression models, random forest, SVM, clustering to identify Volume using different packages in python.
* Performed data visualization with Tableau and generated dashboards to present the findings
* Created and designed reports that will use gathered metrics to infer and draw logical conclusions from past and future behavior.
* Worked independently and collaboratively throughout the project lifecycle including data extraction/preparation, design and implementation of scalable machine learning analysis and solutions, and documentation of results.
* Performed Classification using supervised algorithms like Logistic Regression, Decision trees, KNN, Naive Bayes.
* Performed data profiling to merge the data from multiple data sources.
* Knowledge of other relational database platforms such as Oracle, NoSQL

**Environment:** Python 3, MySQL, Matplotlib, Seaborn, Linear Regression, Logistic Regression, Random Forest, Support Vector Machines, KNN, Tableau.

|  |
| --- |
| **Client: Insync, San Francisco, CA** |
| **Role: Machine Learning Consultant April 2013 – Feb 2015** |
|  |

**Responsibilities:**

* Worked as data identification, collection, and exploration, cleaning the model, participate in model identification.
* Worked with both supervised and unsupervised data algorithms and evaluated the models, tested and validated before selecting the best fit model for predictions.
* Working with large sets of complex datasets that include structured, semi-structured and unstructured data and discover meaningful business insights.
* Utilized Machine Learning Algorithms such as Linear Regression, Logistic Regression, Naive Bayes, Random Forests, K-means, & KNN for data analysis.
* Skilled in Advanced Regression Modelling, Correlation, Multivariate Analysis, Model Building, Business Intelligence tools and application of Statistical Concepts.
* Enforced F-Score, AUC/ROC, Confusion Matrix, Precision, and Recall evaluating different model's performance.
* Completed many tasks from collecting the data and exploring the data and interpreting the statistical information.
* Identifying the data needs and requirements and work with other members of the IT- organization to deliver proper Data Visualization and reporting solutions to those needs.
* Customer segmentation based on their behavior or specific characteristics like age, region, income, geographical location and applying Clustering algorithms to group the customers based on their similar behavior patterns.
* Designed predictive models using the Machine Learning platform. Worked with numerous data visualization tools in python like matplotlib, seaborn, ggplot, pygal.

**Environment:** Python 3, NumPy, Matplotlib, Jupyter notebook Tableau9.3, logistic regression, random forest, neural networks, SVM, JSON, XML, MLLib, Tensorflow.

**Education:**

Bachelors in Information & Communication Technology

**References:** Available upon request.