**Sundar Pandey**  
[pandeysundar005@gmail.com](mailto:pandeysundar005@gmail.com)

(903) 326-3772

**PROFESSIONAL SUMMARY**

* 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 **Python**.
* Experience in dimensionality reduction using techniques like **PCA and LDA**.
* Experience in data analytics, predictive analysis like Classification, Regression, Recommender Systems.
* 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**.
* Expertise and knowledge in **TensorFlow** to do machine learning/deep learning package in **python**.
* Good Knowledge on **Natural Language Processing (NLP)** and **Time Series Analysis** and Forecasting using **ARIMA model** in **Python**.
* Good knowledge in **Tableau, Power BI** for interactive data visualizations.
* Exposure with AWS cloud services **EC2**, S3, IAM, **AWS security and Encryption**

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **Languages** | Python |
| **Packages** | json, plyr, pandas, numPy, Seaborn, sciPy, matplotlib, sci-kit-learn, Beautiful Soup, |
| **Databases** | SQL, pySpark, Databases SQL-Server, My SQL |
| **Reporting Tools** | MS Office (Word/Excel/Power Point/), Tableau |
| **Version Control Tools** | GitHub |
| **BI Tools** | Tableau, Tableau Server, Tableau Reader |
| **Operating System** | Windows, Linux (Ubuntu, CentOS), Macintosh HD |

### **Education:**

Ph.D. in Sciences: Biological Sciences, 2018 (Miami, FL, USA)

Bachelor’s in technology: Biotechnology, 2007 (Kathmandu, Nepal)

### **WORK EXPERIENCE:**

### **Macy’s, Atlanta, GA. Aug 2018 to Present**

**Data Scientist**

This project is about developing an algorithm that accurately predicts the demand of products among multiple classes based on the historical sales data available on multiple products. Further, the aim was to improve the profit by maintaining the right stock of products whose demand is high while avoiding the loss of maintaining unnecessary products.

* Collaborated with the business analyst on the requirements of the project and explored the data from the database querying (**SQL**) search techniques, web services etc.
* Preparing data using techniques like dimensionality reduction for reduction of features using (**PCA, t-SNE**), cleaning the data using libraries of Python.
* Applying advanced statistical techniques (Bayesian, sampling and experimental design) while performing machine learning algorithms on the heterogenous data.
* Used advanced analytical tools and programming languages such as Python (**NumPy, pandas, SciPy**) for data analysis.
* Constructed and evaluated various types of datasets by performing machine learning models using algorithms and statistical modeling techniques such as **clustering, classification, regression, decision trees, support vector machines**, anomaly detection, sequential pattern discovery, and text mining from Python libraries (**scikits.learn**).
* Performing the Post pruning techniques in machine learning to reduce the complexity of the final classifier which results in improving the predictive analysis by reducing over fitting, using python libraries(**sklearn**).
* Performing predictive analytics and machine learning algorithms especially supervised (**SVM, Logistic Regression, Boosting**), unsupervised (**K-Means, LDA, EM**) and Reinforcement learning (**Random Forests**) methods.
* Obtained better predictive performance of 81% accuracy using ensemble methods like Bootstrap aggregation (**Bagging**) and Boosting (**Adaboost, Gradient).**
* Build decision tree and random forest based on **Entropy, Information gain** and **Gini Impurity** for split criteria.
* Using regularization techniques to solve the over-fitting problem by reducing loss function either by adding multiple (**LASSO or Ridge**) or by performing cross validation.
* Read the different data formats like API (JSON), XML, CSV, Rich Text Format (.rtf), Open Document Text (. odt), HTML (.html), parquet.
* Deployed Spark Ecosystem includes **Spark SQL, pySpark DataFrames**
* Visualized graphs and reports using **matplotlib, and seaborn** packages in python on datasets for analytical models to know the missing values, outliers, correlation between the features.
* Utilizing **Tableau** visualization software for visualizing the results of the model by transforming data into dashboards that look amazing and are also interactive.
* Creating user stories, sub tasks, epics in **JIRA** for the project. To track the flow of the project used Kanban board throughout different phases of lifecycle.

### **Regions Bank, Birmingham, Al Nov 2016 – July 2018**

### **Data Scientist**

Smart Banking Chat Bot- This is an AI based project which uses several ML algorithms for Natural Language Understanding which identifies intent and entities from user issues and generates dialogue. This project can help Banks to add chatbot in their web-application, so that customer can ask question to chatbot to Banks without visiting to Bank.

* Used **Pandas**, **NumPy, Seaborn, SciPy, Matplotlib, Scikit-learn, and NLTK** in Python for developing various machine learning algorithms.
* Installed and used Caffe NLP Framework.
* Worked on different data formats such as JSON, XML and performed machine learning algorithms in Python.
* 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.
* Implemented and Engineered reinforcement, machine learning algorithms enabled analysis of massive quantities of data.
* Created next generation spam filtering model using Random Forest algorithm, optimizing the hyper parameters, tuning the architecture by minimizing false positives, false negatives and achieved accuracy.
* Used ensemble technique to reduce problems related to over-fitting of the training data using **Bagging** and **Boosting**.
* Handled non-elliptical shapes, time and space requirements using hierarchical clustering.
* Performed classification problems using **SMOTE** oversampling technique. Used **MAML** algorithm approach to perform small number of gradient steps while avoiding over-fitting.
* Used **EM** algorithm to perform data clustering in machine learning and for Hidden Markov Models used the Baum-Welch algorithm in NLP

**Environment:** MySQL, Keras, TensorFlow, AWS, Windows 7, MS Excel, Tableau, Python, JIRA.

**Life Technologies, SFO, CA. May 2015 to Oct 2016**

**Data Analyst**

**Responsibilities:**

* Designed, developed and delivered the development life cycle with detailed level of data analysis using technical tools.
* Knowledge of data life cycle – data acquisition, data quality management, data governance, and metadata management.
* Capable of building, articulating, and presenting new ideas to technical, non-technical, and business communities.
* Used analytics data preparation methods, e.g. data validation, data quality assurance, data transformation.
* Assembled large, complex data sets that meet functional or non-functional business requirements.
* Used Python's **pandas** for file conversion, file counts, column analysis and formatted.
* Designed and created multiple worksheets, analytical reports and Data Visualization Dashboards to help users for identifying Key Performance Indicator along with strategic planning in firm using Tableau as Data Visualization as per the requirements of the end user.
* Used various statistical methods like **Hypothesis Testing, Chi-Square test, Control charts, t-Test, ANOVA**, Correlation Techniques, Statistical Process Control and Descriptive Statistics.
* Used various Python libraries like **seaborn, scikit-learn, SciPy** to visualize, analyze the data for machine learning.
* Developed various Statistical Methods, expertise in Text Analytics, created data visualization, build solutions for Data Mining using **R** and **Python**.
* Created **SQL** Schema such as Functions, Views, Procedures, Sequences, Record Type, Triggers and Object Type, performed in coordination with SQL Developer.
* Implemented data pre-processing such as text preprocessing, data cleaning, noise removal, object standardization and lexicon normalization.

**Environment:** Python, NumPy, Pandas, SciPy, Tableau, Tableau, SQL Server.