**Sundar Pandey**

**PROFESSIONAL SUMMARY**

**6 years of professional expertise data science. Seeking to increase data usage efficiency. Achievements include creating data regression models to forecast company stock prices with 30% more accuracy than historical average and achieved 25% improvement.**

* 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**.
* Experience in migration from heterogeneous sources including **Oracle** to **MS SQL Server**.
* In depth knowledge and hands on experience of **Big Data / Hadoop ecosystem (HDFS, Hive, and Sqoop)**.
* Experience in **Apache Spark, Kafka** for **Big Data Processing & Scala Functional** **programming**.
* Experience in manipulating the large data sets with **R packages** like **tidyr, tidyverse, dplyr reshape, lubridate, Caret** and visualizing the data using lattice and **ggplot2 packages**.
* 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 **Microsoft Azure SQL, Machine Learning** and **HDInsight**.
* Good Exposure on **SAS analytics**.
* Good Exposure in deep learning with Tensor flow in **python**.
* Good Knowledge on **Natural Language Processing (NLP)** and **Time Series Analysis** and Forecasting using **ARIMA model** in **Python and R**.
* Good knowledge in **Tableau, Power BI** for interactive data visualizations.
* Experience with AWS cloud services **EC2**, S2, IAM, CloudFront, Athena, **AWS security and Encryption**

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **Languages** | Java 8, Python, R |
| **Packages** | ggplot2, caret, dplyr, Rweka, gmodels, RCurl, C50, twitter, NLP, Reshape2, json, plyr, pandas, numPy, Seaborn, sciPy, matplot lib, sci-kit-learn, Beautiful Soup, Rpy2. |
| **Web Technologies** | HTML5, DHTML and XML, CSS3, Web Services |
| **Data Modelling Tools** | Erwin r 9.6, 9.5, 9.1, 8.x, Rational Rose, ER/Studio, MS Visio, SAP Power designer |
| **Big Data Technologies** | Hadoop, Hive, HDFS, MapReduce, Pig, Kafka |
| **Databases** | SQL, Hive, Impala, Pig, Spark SQL, Databases SQL-Server, My SQL, MS Access, HDFS, HBase, Teradata, Netezza, MongoDB, Cassandra. |
| **Reporting Tools** | MS Office (Word/Excel/Power Point/ Visio), Tableau, Crystal reports XI, Business Intelligence, SSRS, Business Objects 5.x/ 6.x, Cognos7.0/6.0. |
| **ETL Tools** | Informatica Power Centre, SSIS. |
| **Version Control Tools** | GitHub |
| **BI Tools** | Tableau, Tableau Server, Tableau Reader, SAP Business Objects, OBIEE, QlikView, SAP Business Intelligence, Amazon Redshift, or Azure Data Warehouse |
| **Operating System** | Windows, Linux (Ubuntu, CentOS), Macintosh HD |

### **WORK EXPERIENCE:**

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

**Data Scientist/ Machine Learning Engineer**

* 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.
* Repeating steps as required for improving the Scalability, Reliability and performance of our Streaming Data Pipelines which was built on top of **Spark**.
* Importing and exporting data from various RDBMS such as **MySQL, Oracle**, mainframe into Hadoop Distributed File System (**HDFS**), transform the data in Hadoop **MapReduce**, and then export the data back into an RDBMS using **Sqoop**.
* Working with the Hadoop ecosystem, including **HDFS, MapReduce, Hive, and Spark** for managing data processing and storage for big data applications running in clustered systems.
* Read the different data formats like API (JSON), XML, CSV, Rich Text Format (.rtf), Open Document Text (. odt), HTML (.htm, .html), parquet, Avro.
* Deployed Spark Ecosystem includes **Spark SQL, Spark DataFrames, MLlib, GraphX, Spark Streaming, Spark Core API** increase productivity and can be seamlessly combined to create complex workflows.
* Visualized graphs and reports using **matplotlib, seaborn and panda** 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.

**Environment:** MySQL Workbench 5.7, Python 3.6.3, Jupiter notebook 5.5.0, Apache Spark 2.2, Tableau 10.4, Hive 2.3.0, Sqoop 1.4.6, Hadoop 2.7.1, Spark 2.2, JIRA

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

### **Data Scientist**

* Worked with several R packages including **knit, dplyr, Spark, R, Causal Infer, Space-Time.**
* Coded R functions to interface with **Caffe Deep** Learning Framework.
* 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.
* Setup storage and data analysis tools in Amazon Web Services cloud computing infrastructure.
* Implemented end-to-end systems for Data Analytics, Data Automation and integrated with custom visualization tools using R, Mahout, Hadoop and MongoDB.
* Worked as Data Architects and IT Architects to understand the movement of data and its storage and ERStudio9.7.
* 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.

**Environment:** MySQL, Cassandra, Netezza, Linux, SQL, Hadoop, SPSS, Spark, Keras, TensorFlow, AWS, Pig, Hive, Windows 7, MS Excel, Tableau, VBA, R, Shiny, Python, Google Cloud, IBM Jazz, JIRA.

**Ryan LLC, Cleveland, Ohio May 2015 to Oct 2016**

**Data Scientist**

* Implemented and Engineered reinforcement, machine learning algorithms enabled analysis of massive quantities of data.
* Integrated with a diverse team to deliver in an agile methodology for every sprint meeting and made ready for review the specific work.
* Investigated and solved exciting challenges in classification, deep learning, image recognition, and content analysis.
* Proficient with programming languages like **R** and **Python** and 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.
* Solved k-medoids problem using **PAM** and **CLARANS** algorithms. Clustered categorical data using **ROCK** by neighbor and link analysis, **LIMBO** and **COOLCAT** using information theoretic tools.
* Used **EM** algorithm to perform data clustering in machine learning and for Hidden Markov Models used the Baum-Welch algorithm in NLP.
* Handled clusters of different sizes and shapes using **DBSCAN** algorithm which was also resistant to noise and used for determining **Eps** and **MinPts**.
* Performed classification problems using **SMOTE** oversampling technique. Used **MAML** algorithm approach to perform small number of gradient steps while avoiding over-fitting.
* Ability to debug production code and deliver the results on time. Performed architect, test, and tune and deploy algorithms into production platform.
* Implemented user profile system based on behaviors of user and applied the recommendation systems algorithm based on profiles of users.
* Performed predictive analysis using **scikit-learn, HBase**. Hands-on implementation and coding in Scala.
* Used “big data” technologies including **Map Reduce, Kafka, HBase, Pig, Cassandra**. Proficient with website analytics tool like Google Analytics.

**Environment:** Python 3.4.0, RStudio 0.98.493, Apache HBase, Pig v0.13.0, Classification, Clustering, Regression, NLP, Deep Learning, Reinforcement learning.

**Life Technologies, SFO, CA. May 2013 to Apr 2015**

**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**.
* Analyzed large, complex, multi-dimensional datasets using a variety of tools like **R** and **SPSS.**
* Hands-on knowledge of the Big Data technologies – **Hadoop, Hive, Impala, HBase, Sqoop, Pig, Kafka** etc.
* Created **SQL** Schema such as Functions, Views, Procedures, Sequences, Record Type, Triggers and Object Type, performed in coordination with SQL Developer.
* Build the infrastructure required for optimal extraction, transformation, and loading of data from a wide variety of data sources using SQL and AWS ‘big data’ technologies.
* Implemented data pre-processing such as text preprocessing, data cleaning, noise removal, object standardization and lexicon normalization.
* Performed data ingestion into Big Data platform from distinct data sources using **Sqoop**.
* Coordinated with the developer’s team to develop, implement and design solutions for BI cases to determine Sale KPIs at macro and micro level.

**Environment:** Python, Hadoop, Hive, NumPy, Pandas, SciPy, Map Reduce, Tableau, Sqoop, HBase, HDFS, Tableau, SQL Server.