**[DATA SCIENTIST/ANALYST]**

**Sukhjinder Singh**

•Bellevue (WA) • sukhmailbox@gmail.com • https://www.linkedin.com/in/sukhjindersingh93

**PROFESSIONAL SUMMARY:**

* Data Science Professional with 6+ years of experience in Acquisition of correct Datasets, Data Scrubbing to mine the target data, Data Engineering to extract features utilizing Statistical Techniques, Exploratory Data Analysis with an inquisitive mind, build diverse Machine Learning Algorithms for developing Predictive Models and design Stunning Visualizations to help the growth of Business Profitability.
* Outstanding preeminence in Data extraction, Data cleaning, Data Loading, Statistical Data Analysis, Exploratory Data Analysis, Data Wrangling, Predictive Modeling using R, Python and Data visualization using Power BI.
* Profound knowledge in Machine Learning Algorithms like Linear, Non - linear and Logistic Regression, SVR, Reinforcement Learning, Natural Language Processing, Fuzzy Logic, Random forests, Ensemble Methods, Decision tree, Gradient-Boosting, K-NN, SVM, Naïve Bayes, Clustering (K-means), Deep Learning.
* Proficient in Machine Learning algorithms and Predictive Modeling including Linear Regression, Logistic Regression, Naive Bayes, Decision Tree, Neural Networks, Random Forest, Ensemble Models, SVM, KNN and K-means clustering.
* Solid experience in Deep Learning techniques with Convolutional Neural Networks (CNN), Recursive Neural Networks (RNN), max pooling, normalization and different architectures such as Alexnet, VGG and Darknet.
* Excellent proficiency in model validation and optimization with Model selection, Parameter tuning and K-fold cross validation.
* Strong skills in Statistics Methodologies such as Hypothesis Testing, Principle Component Analysis (PCA).
* Proficient with Python 3.x including Numpy, Scikit-learn, Pandas, Matplotlib and Seaborn.
* Proficient at data visualization tools such as Tableau, Python Matplotlib and Seaborn.\
* Extensive experience in RDBMS such as SQL server 2012, Oracle 9i/10g.
* Explicitly fashioned in writing SQL queries for various RDBMS such as SQL Server, MySQL, Microsoft SQL, Postgre SQL, Teradata and Oracle , NoSQL databases such as MongoDB, HBase and Cassandra to handle unstructured data.
* Solid understanding of RDBMS database concepts including Normalization and master in creating database objects such as tables, views, stored procedures, triggers, row-level audit tables, cursors, indexes and user-defined data-types
* Smart in investigating and scrutinizing large databases like Microsoft Azure, MongoDB , Teradata, Cassandra , Oracle, SQL Server, DB2 .
* Rapidly evaluated Deep Learning frameworks, tools, techniques and approaches for Implementing and building data ingestion pipelines for Neural Networks that included CNNs and RNNs like LSTMs using TensorFlow and Keras.
* Fashioned on different libraries related to Data science and Machine learning like Scikit-learn, OpenCV, NumPy, SciPy, Matplotlib, pandas, Seaborn, Bokeh, nltk, genism, networkx, Stats models, TensorFlow, Theano and Keras.
* Masterly skilled in variant libraries of R such as ggplot2, caret, dplyr, CA Tools, Amelia, Beautiful Soup, e1071, lubridate, missForest, caret, CORElearn, BigRF, e1071, rpart, PROC, igraph, tree, nnet, randomForest, LTSA, LSMeans, ROCR, gbm, Rweka, arules, sqldf, RODBC, RMarkdown.
* Excellence in handling Big Data Ecosystems like Apache Hadoop, MapReduce, Spark, HDFS Architecture, Cassandra, HBase, Sqoop, Hive , Pig, MLlib, ELT.
* Increased customer visibility by developing real-time insights to sales people and sales managers using Tableau, SAS, QlikView, Microsoft BI, Matplotlib, ggplot2, Bokeh, Shiny, Dygraphs resulting in boosting the revenue by 10%.
* Expertise in transforming business requirements into analytical models, designing algorithms, building models, developing data mining & reporting solutions that scales across massive volume of structured and unstructured data.
* Worked on ETL and Report testing using different Data warehousing tools like SSIS, Databricks, AWS, Cognos, Data Mart Worked on AWS RDS, Redshift & Tableau.
* Super-eminent understanding of AWS (Amazon Web Services) , S3, Amazon RDS, Apache Spark RDD , process and concepts. Developing Logical Data Architecture with adherence to Enterprise Architecture.
* Meticulously experienced working on Data quality tools Informatica IDQ (9.1), Informatica MDM (10.1).
* Experience with Data Analytics, Data Reporting, Ad-hoc Reporting, Graphs, Scales, PivotTables and OLAP reporting.
* Progressive involvement in Software Development Life Cycle (SDLC), GIT , Agile methodology and SCRUM process. Strong business sense and abilities to communicate data insights to both technical and nontechnical clients.
* Conscientiously skilled in System Analysis, E-R/Dimensional Data Modeling, Design and implementing RDBMS specific features. Conventionally accessing JIRA tool and other internal issue trackers for the Project developments.
* Robust participation for functioning in fast-paced multi-tasking environment both independently and in the collaborative team. Adequate with challenging projects and work in ambiguity to solve complex problems. A self-motivated exuberant learner.

**TECHNICAL SKILLS:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|

|  |  |
| --- | --- |
| Databases | SQL Server, MS Access, AWS RDS, Teradata, Oracle 9i/10g/11g/12c, MySQL 5.5/5.6, Microsoft SQL 2008/12/14/16 , Postgre SQL |
| NoSQL Databases | MongoDB 3.x, Hadoop HBase 0.98 and Apache Cassandra |
| Programming Languages | Python 2.x/3.x, R 3.x, SQL, Scala, pig C, C++, MATLAB, Java, JavaScript |
| Cloud Technologies | AWS, Microsoft Azure, Open stack, Docker |
| Querying Languages | SQL, NO SQL, PostgreSQL, MySQL, Microsoft SQL |
| Markup languages | XML, HTML, DHTML, XSLT, X Path, X Query and UML |
| Deployment Tools | Anaconda Enterprise v5, Databricks, R-Studio, Azure Machine Learning Studio, Oozie 4.2, AWS Lambda |
| ETL Tools | Informatica Power Center, SSIS, SAS Data Management, Talend |
| Data Modeling Tools | MS Visio, Rational Rose, Erwin r9.6/9.5, ER/Studio 9.7, Star-Schema Modeling |
| Testing Tools | JIRA, HP Quality Center ALM, Testrail |
| Scalable Data Tools | Hadoop, Hive 1.x/2.x, Apache Spark 2.x, Pig 0.15, Map Reduce, Sqoop |
| Operating Systems | Red Hat Linux, Unix, Ubuntu, Debian, Centos, Windows, macOS |
| Reporting & Visualization | Tableau 9.x/10.x, Matplotlib, Seaborn, Bokeh, ggplot, iplots, SAP Business Objects, Tibco Spotfire, Crystal Reports, SSRS, Cognos, Shiny, Microsoft BI, Qlik View |

 |

 **Links:**

* GitHub: sukhjindermultan

**Education:**

* Master’s in Data Science from Washington State University May 2019 | Pullman, WA
* BS in Computer Science from Lovely Professional University May 2015 | Phagwara, India.

**Library Packages:**

• Keras,Pandas,NumPy

• MapReduce,Hive,Hadoop

• Spark, PySpark,

• XGBoost

• Matplotlib

• Scikit-Learn

• Google Analytics

• Bokeh

• Seaborn, ggplot, Dash-Plotly

• Elastic E

**Certifications:**

* Machine Learning
* Statistical Modeling
* Machine Learning with
* Tensor flow On GCP
* AWS certification
* Applied Data Analyst

**Professional Experience:**

**MICROSOFT | REDMOND, WA 11/2019 - Present**

**Data scientist/Data analyst**

**Responsibilities:**

* Championed the design & execution of machine learning projects to address specific business problems determined by consultation with business partners. Exercised Machine Learning Algorithms such as linear regression, SVM, Multivariate Regression, Fuzzy Logic, Naive Bayes, Random Forests, K-means, & KNN for data analysis.
* Worked with data-sets of varying degrees of size and complexity including both structured and unstructured data. Piping and processing massive data-streams in distributed computing environments such as Hadoop to facilitate analysis (ETL).
* Created ETL packages to transform data into the right format and join tables together to get all features required by using SSIS.
* Established the backend server using Flask to crawl audiences' public information from Facebook and apply custom analysis using Watson Personality Insights tool.
* Superintended usage of Python NumPy, SciPy, Pandas, Matplot, Stats packages to perform dataset manipulation, data mapping, data cleansing and feature engineering. Built and analyzed datasets using R and Python.
* Used Data Quality validation techniques to validate Critical Data elements (CDE) and identified many anomalies. Extensively worked on statistical analysis tools and adept at writing code in Advanced Excel, R and Python.
* Enforced model Validation using test and Validation sets via K- fold cross validation , statistical significance testing
* Multi-layers Neural Networks built in Python Scikit-learn, Theano, TensorFlow and keras packages to implement machine learning models and export them into protobuf and performed integration job with client's application.
* Predictive analytics and machine learning algorithms to forecast key metrics in the form of designed dashboards on to AWS (S3/EC2 CLOUD PLATFORMS) and Django platform for the company's core business.
* Extracting the data from Azure Data Lake into HDInsight Cluster (INTELLIGENCE + ANALYTICS) and applying spark transformations & Actions and loading into HDFS.
* Built predictive models including linear regression, Random Forest Regression and Support Vector Regression to predict the claim spending by using python scikit-learn.
* Used GridSearchCV to evaluate each model and to find best parameters for each model.
* Created a report by using Tableau to show client how prediction can help the business.
* R programming language for graphically critiquing the data and performed data mining. Interpreting Business requirements, data mapping specifications and visualized data as per the business requirements using R shiny.
* Fronted the migration of analytics database from Redshift to Databricks and improved the system productivity by 30%.
* Predictive analytics and machine learning algorithms to forecast key metrics in the form of designed dashboards on to Azure and Django platform for the company's core business.
* Predominant practice of Python Matplotlib package and Power BI to visualize and graphically analyses the data. Data pre-processing, Splitting the identified data set into set and Test set using other libraries in python.
* Elucidating the continuous improvement opportunities of current predictive modeling algorithms. Proactively collaborates with business partners to determine identified population segments and develop actionable plans to enable the identification of patterns related to quality, use, cost and other variables.
* Utilized Informatica toolset (Informatica Data Explorer and Data Quality) to inspect legacy data for data profiling.
* Conceptualized and built data models, tools, custom visualizations and dashboards in Tableauthat communicate results to clients, developed a compelling story with the data and optimized their performance.
* Generating weekly, monthly reports for various business users according to the business requirements.

**Environment:**Informatica Power Center R, Python, MATLAB, ETL, Sypder 3.6, Azure ML, HP ALM, Agile, Data Quality, R Studio, Tableau, Data Governance, Supervised & Unsupervised Learning, Java, NumPy, SciPy, h2o, Pandas, SQL Server 2014, AWS (EC2, RDS, S3), Matplotlib, Scikit-learn, HTML, XML, Shiny

**GE GRID SOLUTIONS | REDMOND, WA 05/2019 - 11/2019**

**Data Scientist/Data Analyst**

**Responsibilities:**

* Collaborated with data engineers and operation team to implement ETL process, wrote and optimized SQL queries to perform data extraction to fit the analytical requirements.
* Supervised data collection and reporting. Ensured relevant data is collected at designated stages, entered into appropriate database(s) and reported appropriately. Monitored assignments to assure distribution of workload and the assessment of collection efforts.
* Evaluated the performance of Databricks environment by converting complex Redshift scripts to spark SQL as part of new technology adaption project.
* Data analysis by using Hive to retrieve the data from Hadoop cluster, SQL to retrieve data from RedShift about retail consumer behavior and report on findings to both internal teams and True Fit partner brands and retailers.
* Dealt with missing values by column dropping, row dropping and replacing the missing data with statistical results.
* Converted unstructured records data to structured dataset using NLP techniques and feature engineering.
* Explored and visualized the data to get descriptive statistics and inferential statistics for better understanding the dataset.
* Built predictive models including support Vector Machine, Decision tree, Naive Bayes Classifier, Neural Network to predict whether the thyroid cancer cell is under potential danger of spreading by using python scikit-learn.
* Implemented process using cross-validation and evaluated the result based on different performance matrices.
* Collected feedback and retrained the model to improve the performance.
* Explored and analyzed the customer specific features by using Matplotlib and ggplot2. Extracted structured data from MySQL databases or CRM systems, developing basic visualizations or analyzing A/B test results.
* Built a website application for learning language, a login module, and a questionnaire based on background scripts using flask framework.
* Conventionally designed Fuzzy Logics and implemented statistical tests including Hypothesis testing, AVOVA, Chi-square test to verify models' significance by using R.
* Participated in features engineering such as feature generating, PCA, feature normalization and label encoding with Scikit-learn preprocessing. Data Imputation using variant methods in Scikit-learn package in Python.
* Work with business stakeholders to refine and respond to their ad hoc requests and improve their existing reporting and dashboards as necessary.
* Experimented and chiefly built predictive models including ensemble models using machine learning algorithms such as Logistic regression, Random Forests and KNN to predict customer churn.
* Conducted analysis on customer behaviors and discover value with RMF analysis; applied customer segmentation with clustering algorithms such as K-Means Clustering. Gaussian Mixture Model and Hierarchical Clustering.
* Enforced F-Score, AUC/ROC, Confusion Matrix, Precision, and Recall evaluating different models performance.
* Unveiled and scrutinized new insights by diving deeply into the data combining various sources (database and web-analytics data sources) from across the business, to support decisions and resolve strategic questions
* Designed and implemented a recommendation system which leveraged Google Analytics data and the machine learning models and utilized Collaborative filtering techniques to recommend courses for different customers.
* 10+ dashboards in Tableau for sales managers with instant access to personalized analytics portal, so they can access key business metrics such as time to close opportunity, delay-to-contract, resulting in increased customer satisfaction and improving client’s standing in the Sales Performance management industry.
* Organized reports, an app demo, produced rich data visualizations to model data into human-readable form with the Tableau and Matplotlib to show client how prediction can help the business.

**Environment:**RedShift, Hadoop, HDFS, Python 3.x, R (ggplot2/ caret/ trees/ arules), Tableau (9.x/10.x), Machine Learning (Logistic regression/ Random Forests/ KNN/ K-Means Clustering/ Gaussian Mixture Model / Hierarchical Clustering/ Ensemble methods/ Collaborative filtering), JIRA, GitHub, Agile/ SCRUM.

**WASHINGTON STATE UNIVERSITY - PULLMAN, WA 09/2018­­­ - 05/2019**

**Data Analyst**

**Responsibilities:**

* Involved in all the phases of the project, communicated with the business and prepared Technical Specification Document.
* Planned monthly and quarterly business monitoring reports by creating Excel Pivot summary reports to include System Calendars.
* Transferred the MS Excel Sheet Reports into SSRS based reports by migration the data by using SSIS packages and then Use views and tables and stored procedures to develop new reports.
* Assisted budget coordinator with preparing different community programs, allocating budget, preparing daily financial transactions and allocation reports using Microsoft Excel.
* Reconciled daily financial reports using pivot tables and VLOOKUP.
* Extensively use Python's multiple data science packages like Pandas, NumPy, matplotlib, Seaborn, SciPy, Scikit-learn and NLTK.
* Work on data that was a combination of unstructured and structured data from multiple sources and automate the cleaning using Python scripts.
* Improved fraud prediction performance by using random forest and gradient boosting for feature selection with Python Scikit-learn.
* Implemented machine learning model (logistic regression, XGboost) with Python Scikit- learn.
* Optimize algorithm with stochastic gradient descent algorithm Fine-tuned the algorithm parameter with manual tuning and automated tuning such as Bayesian Optimization.
* Broadly designed easy to follow visualizations using Tableau software and published dashboards on web and desktop platforms.
* Provide day-to-day end-user support and assist users with proactive best practices to enhance and increase their knowledge of Salesforce.
* Maintain a holistic view of all business processes and users in the system to understand cross-functional impacts regarding configuration, process and workflow.
* Executed quantitative analyses that translate data into actionable insights therefore provides analytical and data-driven decision-making support for key projects.
* Documented, organized, designed, built, tested and deployed enhancements to Salesforce custom objects, page layouts, workflows, alerts, reports and complex dashboards within Salesforce.
* Compiled and analyzed sales and marketing reports using SQL and Statistical models.
* Involved in analysis, design and documenting business requirement specifications to build data warehousing extraction programs, end-user reports and queries.
* Worked closely with Associates to find the problems and getting solutions on the tool.

**Environment:**Tableau, MS Access, Salesforce, MS Excel, MS Visio, UML diagrams, Mainframes, SQL Server.

 **MULTICARE HEALTH SYSTEM | REDMOND, WA 08/2017 - 05/2018**

**Data Analyst**

**Responsibilities**:

* Communicated and coordinated with other departments to gather business requirements.
* Gathering all the data that is required from multiple data sources and creating datasets that will be used in analysis. Experience in translating/mapping relational data models into Data Base Schemas.
* Capable of facilitating data discovery sessions involving business subject matter experts.
* Ability to create and maintain conceptual/business, logical and physical data models.
* Performed Exploratory Data Analysis and Data Visualizations using R, and Python.
* In Preprocessing phase, used Pandas and Scikit-Learn to remove or impute missing values, detect outliers, scale features, and applied feature selection (filtering) to eliminate irrelevant features.
* Conducted Exploratory Data Analysis using Python Matplotlib and Seaborn to identify underlying patterns and correlation between features.
* Used Python (NumPy, SciPy, Pandas, Scikit-Learn, Seaborn to develop variety of models and algorithms for analytic purposes.
* Setup storage and data analysis tools in Amazon Web Services cloud computing infrastructure.

 **Environment:** Python, Logistic Regression, clustering, NumPy, Scikit Learn, Map Reduce, Pig, Hive, Pandas, Seaborn, Random Forest and Tableau

**AMBIENTE TECHNOLOGIES -MOHALI, INDIA 01/2015 - 08/2017**

**Associate Data Analyst**

**Responsibilities:**

* Created Database in Microsoft Access by using blank database and created tables and entered dataset manually and Data Types, performed ER Diagram and Basic SQL Queries on that database.
* Microsoft Excel used for formatting data as a table, visualization and analyzing data by using certain methods like **Conditional Formatting, Remove Duplicates, Pivot and Unpivot tables, Created Charts and Sort and Filter Data Set.**
* Applied concepts of probability, distribution and statistical inference on given dataset to unearth interesting findings through use of comparison**, T-test, F-test, R-squared, P-value etc.**
* Performed Statistical Analysis and Hypothesis Testing in Excel by using Data Analysis Tool.
* Interacted with the other departments to understand and identify data needs and requirements and work with other members of the IT organization to deliver data visualization and reporting solutions to address those needs.
* Creating customized business reports and sharing insights to the management.
* Presented a Dashboard for better understanding of dataset to the entire stake Holders.
* Performed module specific configuration duties for implemented applications to include establishing role-based responsibilities for user access, administration, maintenance, and support.
* Worked closely with internal business units to understand business drivers and objectives which can be enabled through effective application deployment.

**Environment**: Microsoft Project 2000, MS Excel, MS PowerPoint, MS Visio and MS Outlook, JIRA, SQL Server 2012/2014. SharePoint2010