**Abhishek Singhal**

**Data Scientist**

**Email:abhi19singhl@gmail.com**

**Phone no: 510-338-9120**

**PROFESSIONAL SUMMARY:**

* Professional qualified **Data Scientist/Data Analyst** with **over 8+ years** of experience in **Data Science** and **Analytics** including Artificial Intelligence/Deep Learning/Machine Learning, Data Mining and Statistical Analysis
* Involved in the entire data science project life cycle and actively involved in all the phases including **data extraction, data cleaning, statistical modeling** and **data visualization** with large data sets of structured and unstructured data, created ER diagrams and schema.
* Experienced with machine learning algorithm such as logistic regression, **random forest, XGboost, KNN, SVM, neural network, linear regression, lasso regression** and k - **means**
* Implemented **Bagging** and **Boosting** to enhance the model performance.
* Strong skills in statistical methodologies such **as A/B** test, experiment design, **hypothesis** test, **ANOVA**
* Extensively worked on **Python 3.5/2.7** (Numpy, Pandas, Matplotlib, NLTK and Scikit-learn)
* Experience in implementing data analysis with various analytic tools, such as **Anaconda 4.0 Jupiter Notebook 4.X, R 3.0** (ggplot2, Caret, dplyr) and **Excel …**
* Solid ability to write and optimize diverse SQL queries, working knowledge of **RDBMS** like **SQL Server 2008**, **NoSql** databases like **MongoDB 3.2**
* Developed API libraries and coded business logic using C#, XML and designed web pages using .NET framework, C#, Python, Django, HTML, AJAX
* Strong experience in **Big Data** technologies like **Spark 1.6, Spark sql, pySpark, Hadoop 2.X, HDFS, Hive 1.X**
* Experience in visualization tools like, **Tableau 9.X, 10.X** for creating dashboards
* Excellent understanding **Agile** and Scrum development methodology
* Used the version control tools like **Git 2.X**and build tools like **Apache Maven/Ant**
* Ability to maintain a fun, casual, professional and productive team atmosphere
* Experienced the full software life cycle in **SDLC**, **Agile, Devops** and Scrum methodologies including creating requirements, test plans.
* Skilled in Advanced Regression Modeling, Correlation, Multivariate Analysis, Model Building, Business Intelligence tools and application of Statistical Concepts.
* Proficient in Predictive Modeling, Data Mining Methods, Factor Analysis, **ANOVA**, **Hypothetical testing**, normal distribution and other advanced statistical and econometric techniques.
* Developed predictive models using Decision **Tree**, **Random Forest**, **Naïve Bayes**, **Logistic Regression**, **Social Network Analysis**, **Cluster Analysis**, and Neural Networks.
* Experienced in Machine Learning and Statistical Analysis with **Python Scikit**-Learn.
* Experienced in **Python** to manipulate data for data loading and extraction and worked with **python** libraries like **Matplotlib**, **Numpy**, **Scipy** and **Pandas** for **data analysis**.
* Worked with complex applications such as **R**, **SAS**, **Matlab** and **SPSS** to develop neural network, cluster analysis.
* Strong **C#, SQL** programming skills, with experience in working with functions, packages and triggers.
* Skilled in performing **data parsing**, data ingestion, data manipulation, **data architecture**, **data modelling** and data preparation with methods including describe data contents, compute descriptive statistics of **data, regex, split and combine, Remap, merge, subset, reindex, melt and reshape**.
* Experienced in Visual Basic for Applications and VB programming languages C#, .NET framework to work with developing applications.
* Worked with **NoSQL** Database including **Hbase**, **Cassandra** and **MongoDB**.
* Experienced in **Big Data with Hadoop, HDFS, MapReduce, and Spark**.
* Experienced in Data Integration Validation and Data Quality controls for **ETL** process and **Data** Warehousing using **MS Visual Studio SSIS, SSAS, SSRS**.
* Proficient in **Tableau** and **R**-**Shiny** data visualization tools to analyze and obtain insights into large datasets, create visually powerful and actionable interactive reports and dashboards.
* Automated recurring reports using **SQL** and Python and visualized them on **BI** platform like **Tableau**.
* Worked in development environment like **Git** and **VM**.
* Excellent communication skills. Successfully working in fast-paced multitasking environment both independently and in collaborative team, a self-motivated enthusiastic learner.

## EDUCATION:

Bachelor of Computer Science.

## TOOLS AND TECHNOLOGIES:

|  |  |
| --- | --- |
| **Data Modeling Tools** | Erwin r9.6, 9.5, 9.1, 8.x, Rational Rose, ER/Studio, MS Visio, SAP Power designer. |
| **Programming Languages** | Oracle PL/SQL, Python, SQL, T-SQL, UNIX shell scripting, Java, Apache Spark, MATLAB, R, Big Data(Hive, Pig, MapReduce), Unix, MPI, HTML, AWS |
| **Scripting Languages** | Python (NumPy, SciPy, Pandas, Gensim, Keras), R (Caret, Weka, ggplot) |
| **Big Data Technologies** | Hadoop, Hive, HDFS, MapReduce, Pig, Kafka. |
| **Reporting Tools** | Crystal reports XI, Business Intelligence, SSRS, Business Objects 5.x/ 6.x, Cognos7.0/6.0, Tableau. |
| **ETL** | Informatica Power Centre, SSIS. |
| **Project Execution Methodologies** | Ralph Kimball and Bill Inmon data warehousing methodology, Rational Unified Process (RUP), Rapid Application Development (RAD), Joint Application Development (JAD) |
| **BI Tools** | Tableau, Tableau server, Tableau Reader, SAP Business Objects, OBIEE, Qlikview, SAP Business Intelligence, Amazon Redshift, or Azure Data Warehouse |
| **Tools** | MS-Office suite (Word, Excel, MS Project and Outlook), Spark MLlib, Scala NLP, MariaDB, Azure, SAS. |
| **Databases** | Oracle, Teradata, Netezza, Microsoft SQL Server, MongoDB, HBase, Cassandra. |
| **Operating Systems** | Windows, UNIX, MS DOS, Sun Solaris. |
| **Data Modeling Tools** | Erwin […] Sybase Power Designer, ER Studio, Enterprise Architect, Oracle Designer, MS Visio. |

**PROFESSIONAL EXPERIENCE:**

|  |  |
| --- | --- |
| **Client: JB Hunt Lowell, Arkansas.** | **Aug 2018 - Till Date** |
| **Role: Data scientist** |  |

**Description:** J.B. Hunt Transport Services, Inc. is a trucking and transportation company that was founded by Johnnie Bryan Hunt, and based in the Northwest Arkansas city of Lowell.

**Responsibilities:**

* Performed Data Profiling to learn about behavior with various features of USMLE examinations of various student patterns.
* Evaluated models using Cross Validation, Log loss function, **ROC** curves and used **AUC** for feature selection and elastic technologies like Elastic Search, Kibana etc
* Addressed over fitting by implementing of the algorithm regularization methods like **L2** and **L1**.
* Implemented statistical modeling with **XGBoost** machine learning software package using Python to determine the predicted probabilities of each model.
* Created master data for modeling by combining various tables and derived fields from client data and students LORs, essays and various performance metrics.
* Formulated a basis for variable selection and **Grid Search, KFold** for optimal **hyper parameters**
* Utilized Boosting algorithms to build a model for predictive analysis of student’s behaviour who took USMLE exam apply for residency.
* Used **numpy, scipy, pandas, nltk(Natural Language Processing Toolkit), matplotlib** to build the model.
* Formulated several graphs to show the performance of the students by demographics and their mean score in different USMLE exams.
* Application of various **Artificial Intelligence(AI)/**machine learning algorithms and statistical modeling like decision trees, text analytics, **natural language processing(NLP), supervised and unsupervised**, **regression models .**
* Used **Principal Component Analysis** in feature engineering to analyze high dimensional data.
* Performed Data Cleaning, features scaling, features engineering using pandas and **numpy** packages in python and build models using deep learning frameworks.
* Created deep learning models using **Tensor flow** and **keras** by combining all tests as a single normalized score and predict residency attainment of students.
* Used XGB classifier if the feature is an categorical variable and XGB regressor for continuous variables and combined it using **Feature Union** and **Function Trans fomer** methods of Natural Language Processing.
* Used **Onevs Rest**Classifier to fit each classifier against all other classifiers and used it on multiclass classification problems.
* Implemented application of various machine learning algorithms and statistical modeling like **Decision Tree, Text Analytics, Sentiment Analysis, Naive Bayes, Logistic Regression and Linear Regression** using Python to determine the accuracy rate of each model.
* Created and designed reports that will use gathered metrics to infer and draw logical conclusions of past and future behavior.
* Generated various models by using different machine learning and deep learning frameworks and tuned the best performance model using Signal Hub.
* Created data layers as signals to Signal Hub to predict new unseen data with performance not less than the static model build using deep learning framework.

**Environment:** Python 2.x,3.x, Hive, AWS, Linux, Tableau Desktop, Microsoft Excel, NLP, Deep learning frameworks such as TensorFLow, Keras, Boosting algorithms etc .

|  |  |
| --- | --- |
| **Client: Common securitization solutions Maryland** | **May 2017 - Jul 2018** |
| **Role: Data scientist** |  |

**Description:** Common Securitization Solutions (CSS) is a joint venture owned by Fannie Mae and Freddie Mac and was established to build a universal platform — or Common Securitization Platform (CSP) — for the issuance and management of mortgage-backed securities. The CSP will help Fannie Mae and Freddie Mac provide even greater liquidity to the mortgage finance industry by increasing the efficiency of the securitization process. The entire housing chain, from borrowers to investors, will benefit from what we do for our customers.

**Responsibilities:**

* Performed Data Profiling to learn about behavior with various features such as traffic pattern, location, time, Date and Time etc.
* Application of various Artificial Intelligence(AI)/machine learning algorithms and statistical modeling like decision trees, text analytics, natural language processing(NLP), supervised and unsupervised, **regression models**, **social network analysis**, neural **networks, deep learning**, **SVM**, clustering to identify Volume using scikit-learn package in **python**, **Matlab**.
* Utilized **Spark, Snowflake, Scala, Hadoop, HQL, VQL, oozie, pySpark, Data Lake, Tensor Flow, HBase, Cassandra, Redshift, MongoDB, Kafka, Kinesis, Spark Streaming, Edward, CUDA, MLLib**, AWS, 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.
* Created and connected SQL engine through C# to connect database, developed API libraries and business logic using C#, XML and Python
* Exploring DAG’s, their dependencies and logs using **AirFlow** pipelines for automation
* Performed data cleaning and feature selection using MLlib package in **PySpark**and working with deep learning frameworks such as **Caffe**, **Neon** etc
* Developed **Spark**/**Scala**, **Python, R** for regular expression (**regex**) project in the **Hadoop**/**Hive** environment with **Linux**/**Windows** for big data resources. Used clustering technique **K-Means** to identify outliers and to classify unlabeled data.
* Created user friendly interface for quick view of reports by using C#, JSP, XML and developed expandable menu that show drilldown data on graph click
* Evaluated models using Cross Validation, Log loss function, **ROC** curves and used **AUC** for feature selection and elastic technologies like Elastic Search, Kibana etc
* Categorised comments into positive and negative clusters from different social networking sites using **Sentiment Analysis** and **Text Analytics**
* Tracking operations using sensors until certain criteria is met using **AirFlow** technology.
* Responsible for different Data mapping activities from Source systems to **Teradata** using utilities like TPump, FEXP, MLOAD, BTEQ, FLOAD etc
* Analyze traffic patterns by calculating autocorrelation with different time lags.
* Ensured that the model has low False Positive Rate and Text classification and sentiment analysis for unstructured and semi-structured data.
* Addressed overfitting by implementing of the algorithm regularization methods like **L2** and **L1**.
* Used Principal Component Analysis in feature engineering to analyze high dimensional data.
* Created and designed reports that will use gathered metrics to infer and draw logical conclusions of past and future behavior.
* Performed Multinomial Logistic **Regression**, **Random forest**, **Decision** Tree, **SVM** to classify package is going to deliver on time for the new route.
* Performed data analysis by using **Hive** to retrieve the data from **Hadoop cluster**, **Sql** to retrieve data from **Oracle** database and used **ETL**for data transformation.
* Used **MLlib,** Spark's Machine learning library to build and evaluate different models.
* Implemented rule based expertise system from the results of exploratory analysis and information gathered from the people from different departments.
* Performed Data Cleaning, features scaling, features engineering using pandas and **numpy** packages in python and build models using SAP Predictive Analytics.
* Developed **Map Reduce** pipeline for feature extraction using **Hive**and **Pig**.
* Created Data **Quality Scripts** using **SQL** and **Hive** to validate successful data load and quality of the data. Created various types of data visualizations using Python and **Tableau /Spotfire**.
* Communicated the results with operations team for taking best decisions.
* Collected data needs and requirements by Interacting with the other departments.

**Environment:** Python 2.x, CDH5, HDFS, C#, Hadoop 2.3, Hive, Impala, AWS, Linux, Spark, Tableau Desktop, SQL Server 2012, Microsoft Excel, Matlab, Spark SQL, Pyspark.

|  |  |
| --- | --- |
| **Client: WEX Inc - South Portland, ME** | **Jan 2016 - Apr 2017** |
| **Role: Data Scientist** |  |

**Description:** WEX is a global leader in payments solutions. We simplify the complexities of payment systems across continents and industries—including Fleet, Corporate Payments, and Health.

**Responsibilities:**

* Involved in Design, Development and Support phases of Software Development Life Cycle (**SDLC**)
* Performed data **ETL** by collecting, exporting, merging and massaging data from multiple sources and platforms including **SSRS/SSIS (SQL** Server Integration Services) in **SQL Server.**
* Programming experience with **.NET framework, C#, Visual Studio 2005/2008** to build web based, client/server architecture and to produce reports with **C#** and **JSP**.
* Worked with cross-functional teams (including data engineer team) to extract data and rapidly execute from **MongoDB** through **MongDB** connector for **Hadoop**.
* Performed data cleaning and feature selection using **MLlib** package in **PySpark.**
* Performed partitional clustering into 100 by k-means clustering using Scikit-learn package in **Python** where similar hotels for a search are grouped together.
* Used **Python** to perform **ANOVA** test to analyze the differences among hotel clusters.
* Implemented application of various machine learning algorithms and statistical modeling like Decision Tree, Text Analytics, Sentiment Analysis, Naive Bayes, Logistic Regression and Linear Regression using Python to determine the accuracy rate of each model.
* Determined the most accurately prediction model based on the accuracy rate.
* Used text-mining process of reviews to determine customers' concentrations.
* Delivered analysis support to hotel recommendation and providing an online A/B test.
* Designed Tableau bar graphs, scattered plots, and geographical maps to create detailed level summary reports and dashboards.
* Developed **hybrid** model to improve the accuracy rate.

**Environment:** Python, PySpark, C#, Tableau, MongoDB, Hadoop, SQL Server, SDLC, ETL, SSIS, recommendation systems, Machine Learning Algorithms, text-mining process, A/B test.

|  |  |
| --- | --- |
| **Client: Rabobank - New York, NY** | **Mar 2014 - Dec 2015** |
| **Role: Data Analyst** |  |

**Description:** Rabobank is a Dutch multinational banking and financial services company headquartered in Utrecht, Netherlands. It is a global leader in food and agriculture financing and sustainability-oriented banking.

**Responsibilities:**

* Involved with **Business Analysts** team in requirements gathering and in preparing functional specifications and changing them into technical specifications.
* Used **Erwin** and **Visio** to create **3NF** and dimensional data models and published to the business users and **ETL** / **BI** teams.
* Involved in Data mapping specifications to create and execute detailed system test plans. The data mapping specifies what data will be extracted from an internal data warehouse, transformed and sent to an external entity.
* Managed full **SDLC** processes involving requirements management, workflow analysis, source data analysis, data mapping, metadata management, data quality, testing strategy and maintenance of the model.
* Created and maintained Logical and Physical models for the data mart. Created partitions and indexes for the tables in the **data mart**.
* Multiple phased projects to develop **EDW** (Enterprise Data Warehouse) and **Data Marts** to support Business Intelligence needs as per the requirements of the client.
* Performed data profiling and analysis applied various data cleansing rules designed data standards and architecture/ designed the relational models.
* Gathering, reviewing business requirements and Analyzing data sources from **Excel/SQL** for design. Development, testing, and production rollover of reporting and analysis projects within **Tableau** **Desktop**.
* Wrote complex **SQL** queries for validating the data against different kinds of reports generated by Business Objects.
* Interacting with the Business Users for gathering design requirements and taking feedback on improvements.
* Developed **SQL** queries in **SQL** **Server** management studio, Toad and generated complex reports forth end users.
* Involved in extensive **DATA** validation by writing several complex **SQL** queries and Involved in back-end testing and worked with data quality issues.

**Environment:** Erwin, Visio, 3NF, ETL, EDW, Data marts, SQL, SQL Server, Excel.

|  |  |
| --- | --- |
| **Client: Syntel - Hyderabad, India** | **Dec 2012 - Feb 2014** |
| **Role: Data Analyst** |  |

**Description:** Atos is a global leader in digital transformation with 120,000 employees in 73 countries and annual revenue of € 13 billion. European number one in Cloud, Cybersecurity and High-Performance Computing, the Group provides end-to-end Orchestrated Hybrid Cloud, Big Data, Business Applications and Digital Workplace solutions through its Digital Transformation Factory, as well as transactional services through Worldline, the European leader in the payment industry.

**Responsibilities:**

* Queried data from **SQL** server, imported other formats of data and performed data checking, cleansing, manipulation and reporting using **SAS** (Base and Macro) and **SQL**.
* Built loss forecast models using multitude credit data, census data and insurance data.
* Built claim duration model for worker's comp indemnity loss reserve using survival analysis.
* Researched and developed fraud detection model strategy. Planned and documented strategy in white paper and presentation to management.
* Used extreme value theory and generalized Pareto distribution to fit excess liability loss data.
* Performed ad hoc data analysis and reporting using **SAS** and **Excel**.
* Provide guidance, training and sharing **SAS** programming and predictive modeling methodologies to team members

|  |  |
| --- | --- |
| **Client: People Tech Group - Hyderabad, India** | **Jan 2011 - Nov 2012** |
| **Role: Systems Analyst** |  |

**Description:** People Tech is an emerging leader in the Enterprise Applications and IT Services marketplace. People Tech draws its expertise from strategic partnerships with technology leaders like Microsoft, Oracle, and SAP and combines that with the deep understanding of its employees.

**Responsibilities:**

* Analyze business information requirements and model class diagrams and/or conceptual domain models.
* Gather & Review Customer Information Requirements for **OLAP** and building the data mart.
* Performed document analysis involving creation of Use Cases and Use Case narrations using Microsoft Visio, in order to present the efficiency of the gathered requirements.
* Calculated and analyzed claims data for provider incentive and supplemental benefit analysis using Microsoft Access and **OracleSQL**.
* Analyzed business process workflows and assisted in the development of **ETL** procedures for mapping data from source to target systems.
* Worked with **BTEQ** to submit **SQL** statements, import and export data, and generate reports in **Terra**-data**.**
* Responsible for defining the key identifiers for each **mapping/interface**
* Responsible for defining the functional requirement documents for each source to target interface.
* Document, clarify, and communicate requests for change requests with the requestor and coordinate with the development and testing team.
* Coordinated meetings with vendors to define requirements and system interaction agreement documentation between client and vendor system.
* Enterprise Metadata Library with any changes or updates.
* Document data quality and traceability documents for each source interface.
* Establish standards of procedures.

**Environment:** SQL Server 2008R2/2005 Enterprise, SSRS, SSIS, Crystal Reports, Windows Enterprise Server 2000, DTS, SQL Profiler, and Query Analyzer