Name : XXXXXXXXXXXXXXXXXXX
Ph.No: XXXXXXXXXXXXXXXX
Email ID : XXXXXXXXXXXXXXXX

PROFESSIONAL SUMMARY:

- Professional Qualified **Data Scientist/Data Analyst** with over 7 years of experience in Data science and Analytics including **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.
- Applied transformer and recurrent language models BERT, GPT2, LSTM.
- Worked on general-purpose architectures (BERT, GPT-2, Roberta, XLM, Distil Bert, XLNeT, etc.,) provided by Transformers (pytorch-transformers/pytorch-Pertained-Bert) for Natural Language Understanding (NLU) and Natural Language Generation (NLG).
- Designed the new Insight Lab (Saas & PaaS) product for marketing intelligence insights for digital campaigns in attribution and customer experience for F100 clients.
- Conduct market intelligence to determine market requirements for existing and future planning.
- Experienced with machine learning algorithm such as logistic regression, **random forest, XGboost, KNN, SVM, neural network, linear regression, lasso regression** and k – **means data.**
- Developed Predictive Analytics using PySpark and Spark SQL on Data bricks to extract, transform and uncover insights from the raw data.
- Responsible for estimating the cluster size, monitoring, and troubleshooting of the Spark data bricks cluster.
- Involved in Data ingestion to Azure Data Lake, Azure Data bricks by building pipelines in Azure Data Factory.
- Implemented **Bagging** and **Boosting** to enhance the model performance.
- Involved working on different databases like json, SPARK/HADOOP, XML, NoSQL and SQL of different platforms etc.
- Proficient in Power BI, Tableau, Qlik and R-Shiny data visualization tools to analyze and obtain insights into large datasets and to create visually powerful and actionable interactive reports and dashboards.
- Worked with packages like ggplot2 and shiny in R to understand data and developing applications.
- 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 **Mongo DB 3.2**
- Strong experience in **Big Data** technologies like **Spark 1.6, Sparksql, pySpark, Hadoop 2.X, HDFS, Hive 1.X**
- Experience in visualization tools like, **Tableau 9.X, 10.X** for creating dashboards.
- Handling more than 200 service requests per week for Cherwell and Access Management portal.
- Track and route issues in Help Desk software ticketing system (Cherwell Service Management).
- Excellent understanding **Agile** and Scrum development methodology
- Used the version control tools like **Git 2.X.**
- Hands on experience on Unified Data Analytics with Data bricks, Data bricks Workspace User Interface, Managing Data Bricks Notebooks, Delta Lake with Python, Delta Lake with Spark SQL.
- Passionate about gleaning insightful information from massive data assets and developing a culture of sound, data-driven decision making
- Ability to maintain a fun, casual, professional and productive team atmosphere
- Experienced the full software life cycle in **SDLC**, **Agile** and Scrum methodologies.
- 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**, **Mat lab** and **SPSS** to develop neural network, cluster analysis.
- 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.
- Skilled in performing **data parsing**, data manipulation 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**.
- Strong **SQL** programming skills, with experience in working with functions, packages and triggers.
- Experienced in Visual Basic for Applications and **VB** programming languages to work with developing applications.
- Worked with **NoSQL** Database including **Hbase**, **Cassandra** and **Mongo DB**.
- Experienced in **Big Data with Hadoop, HDFS, Map Reduce, and Spark**.
- Experienced in Data Integration Validation and Data Quality controls for **ETL** process and **Data** Warehousing using **MS Visual Studio SSIS, SSAS, and 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**.

Education Details: Bachelors in Electrical engineer from Pakistan

TECHNICAL SKILLS:

Data Analytics Tools/Programming:	Python (Numpy, Scipy, pandas, Genism, Keras), R (Caret, Weka, ggplot), MATLAB, Microsoft SQL Server, Oracle PLSQL, Python.
Analysis &Modelling Tools	Erwin, Sybase Power Designer, Oracle Designer, Erwin, Rational Rose, ER/Studio, TOAD, MS Visio, & SAS.
Data Visualization	Tableau, Visualization packages, Microsoft Excel.
ETL Tools	Informatica Power Centre, Data Stage, Ab Initio, Talend.
OLAP Tools	MS SQL Analysis Manager, DB2 OLAP, Congo's Power Play.
Languages	SQL, PL/SQL, T-SQL, XML, HTML, UNIX Shell Scripting, C, C++, AWK, JavaScript
Databases:	Oracle, Teradata, DB2 UDB, MS SQL Server, Netezaa, Sybase ASE, Informix, Mongo DB, Hbase, Cassandra, AWS RDS.
Project Execution Methodologies	Ralph Kimball and Bill Inmon data warehousing methodology, Rational Unified Process (RUP), Rapid Application Development (RAD), Joint Application Development (JAD).
Tools & Software	TOAD, MS Office, BTEQ, Teradata SQL Assistant.
Methodologies	Ralph Kimball, COBOL
Reporting Tools	Business Objects XIR2, Congo's Impromptu, Informatica Analytics Delivery Platform, Micro Strategy, SSRS, and Tableau.
Tools	MS Office Suite, Scala, NLP, Maria DB, SAS, Spark MLib Kibana, Elastic search packages, VSS
Languages	SQL, T-SQL, Base SAS and SAS/SQL, HTML, XML.
Operating Systems	Windows, UNIX (Sun-Solaris, HP-UX), Windows NT/XP/Vista, MSDOS.

PROFESSIONAL SUMMARY:

Client: Numerator

Data Scientist / Machine Learning Engineer **Aug 2021 - Till Date**

Responsibilities:

- Gathering, retrieving and organizing data and using it to reach meaningful conclusions.
- Developed a system for **collecting data** and **generating their findings** into reports that improved the company.

- Worked on different data formats such as JSON, XML and performed machine learning algorithms in Python.
- Experience in implementation of advance deep learning algorithms such as Universal Sentence Encoder, BERT for NLP applications such as Semantic Similarity and Question Answering system.
- Implemented USE, BERT advance deep learning algorithms
- Database Design Tools and Data Modeling: Star Schema/Snowflake Schema modeling, Fact & Dimensions tables, physical & logical data modeling, Normalization and De-normalization techniques, Kimball & Inmon Methodologies
- Consulting application teams on the job a bends in Cherwell Service managements and working on it to resolve them in control-m with application teams.
- Monitoring firewall request cues in Cherwell service management to review, and execute firewall requests.
- Fronted the migration of analytics database from Redshift to Data bricks and improved the system productivity by 30%.
- Involved in projects to improve customer sentiment and built models in R Shiny
- Setting up the analytics system to provide insights.
- Initially the data was stored in **Mongo DB**. Later the data was moved to **Elastic search**.
- Used **Kibana** to visualize the data collected from Twitter using Twitter REST APIs.
- Developed a multi class, multi label 2-stage classification model to identify depression- related tweets and classify **depression- indicative symptoms**. Utilized the created model to calculate the severity of depression in a patient using Python, Scikit learn, **Weka** and **Meka**.
- Evaluated the performance of Data bricks environment by converting complex Redshift scripts to spark SQL as part of new technology adaption project.
- Conceptualized and created a knowledge graph database of news events extracted from tweets using Java, Virtuoso, Stanford CoreNLP, Apache Jena, and RDF.
- Producing and maintaining internal and **client-based reports**.
- Creating stories with data that a non-technical team could also understand.
- Worked on **Descriptive, Diagnostic, Predictive and Prescriptive analytics**.
- Implementation of **Character Recognition** using Support vector machine for performance optimization.
- Monitored the Data quality and integrity of data was maintained to ensure effective functioning of department.
- Experience in ETL Process with Snowflake 3.0, Talend open studio 7.0.1, Penton Kettle 7.1, Informatica Power Center 10.2, and SQL Server Integration Services 2017 (SSIS)
- Conduct market intelligence to determine market requirements for existing and future planning.
- Used R and python for Exploratory Data Analysis, A/B testing, HQL, VQL, Data Lake, AWS Redshift, oozie, pySpark, Anova test and Hypothesis test to compare and identify the effectiveness of Creative Campaigns.
- Hands on experience in developing customer scorecards and business dashboards using Shiny in R and AWS Quick Sight.
- Managed database design and implemented a comprehensive Star-Schema with shared dimensions.
- Implemented **Normalization Techniques** and build the tables as per the requirements given by the business users.
- Mining large data sets using **sophisticated analytical techniques** to generate insights and inform business decisions.
- **Building and testing hypothesis**, ensuring statistical significance and building statistical models for business application.
- Developed Machine Learning algorithms with **Spark MLib standalone and Python**.
- Design and develop analytics, machine learning models, and visualizations that drive performance and provide insights, from prototyping to production deployment and product recommendation and allocation planning.
- Performed **data pre-processing tasks** like **merging, sorting, finding outliers, missing value imputation, data normalization**, making it ready for statistical analysis.
- Implemented various machine learning models such as regression, classification, Tree based and Ensemble models.

- Created Machine Learning and statistical methods, (SVM, CRF, HMM, sequential tagging) or willingness to intensely learn.
- Computing A/B testing frameworks, clickstream and time spent databases using Airflow
- Building data platforms for analytics, advanced analytics in **Azure**.
- Managing Tickets using basic SQL queries.
- Designed and implemented end-to-end systems for Data Analytics and Automation, integrating custom visualization tools using R, Tableau, and **Power BI**.

Environment: Python 3.6.4, R Studio, MLib, Regression, A/B Test, SQL Server, Hive, Hadoop Cluster, ETL, Tableau, NumPyPandas, Matplotlib, Power BI, Scikit-Learn, ggplot2, Shiny, Tensor Flow, Teradata.

Client: Alteryx, CA **Apr2020-Jul 2021**
Data Scientist
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.
- Performed data analysis by using Hive to retrieve the data from Hadoop cluster, SQL to retrieve data from Redshift.
- Developed Spark pipelines for data preprocessing, data ingestion algorithms for ingestion into Data bricks environment.
- Experienced using State of the Art Algorithms like BERT and XLNET for text classification.
- Involved in development of Web Services using REST API for sending and getting data from the external interface in the JSON format.
- Worked on Shiny and R application showcasing machine learning for improving the forecast of business.
- Proficiency with various data visualization tools like Tableau, Matplotlib/Seaborn in Python, and ggplot2/Rshiny in R to create interactive, dynamic reports, and dashboards.
- Explored and analyzed the customer specific features by using Spark SQL.
- Performed univariate and multivariate analysis on the data to identify any underlying pattern in the data and associations between the variables.
- Utilized QLIK 10.0 and SNOWFLAKE 3.0 to develop API interfaces to logistical systems and pull from SQL based databases and processed ETL.
- Strong Data Modeling experience in ODS, Dimensional Data Modeling Methodologies Likes Star Schema, Snowflake Schema.
- Involved in Data ingestion to Azure Data Lake, Azure Data bricks by building pipelines in Azure Data Factory.
- Performed data imputation using Scikit-learn package in Python.
- Work experience with Cherwell Service Management tool for tickets.
- Participated in features engineering such as feature intersection generating, feature normalize and label encoding with Scikit-learn preprocessing.
- Used Python 3.X (numpy, Scipy, pandas, Scikit-learn, seaborn) and Spark 2.0 (PySpark, MLib) to develop variety of models and algorithms for analytic purposes.
- Used RMSE score, Confusion matrix, ROC, AUC, Cross validation and A/B testing to evaluate model performance in both simulated environment and real world with high recall rates as high as 94%.
- Implemented Statistical Analysis and Testing including Hypothesis testing, Anova, Survival Analysis, Longitudinal Analysis, Experimental Design, Sample Determination and A/B testing to analyze customer behavior and offer customized products, reduce delinquency rate and default rate from 6% to 3%.
- Developed and implemented predictive models using machine learning algorithms such as linear regression, classification, multivariate regression, Naive Bayes, Random Forests, K-means clustering, KNN, PCA and regularization for data analysis.
- Conducted analysis on assessing customer consuming behaviors and discover value of customers with RMF analysis; applied customer segmentation with clustering algorithms such as K-Means Clustering and Hierarchical Clustering.
- Built regression models include: Lasso, Ridge, SVR, and XGboost to predict Customer Life Time Value.

- Built classification models include: Logistic Regression, SVM, Decision Tree, and Random Forest to predict Customer Churn Rate.
- Used F-Score, AUC/ROC, Confusion Matrix, MAE, and RMSE to evaluate different Model performance.

Environment: s: AWS Redshift, EC2, EMR, Hadoop Framework, S3, HDFS, Spark (PySpark, MLib, Spark SQL), Python 3.x (Scikit-Learn/Scipy/Numpy/Pandas/NLTK/Matplotlib/Seaborn), Tableau Desktop (9.x/10.x), Tableau Server (9.x/10.x), Machine Learning (Regressions, KNN, SVM, Decision Tree, Random Forest, XGboost, LightGBM, Collaborative filtering, Ensemble), NLP, Teradata, Git 2.x, Agile/SCRUM

Client: Resurface Labs, CO **Apr 2019 - Mar 2020**
Data Scientist
Responsibilities:

- Involved in gathering, analyzing and translating business requirements into analytic approaches.
- Worked with Machine learning algorithms like Neural network models, Linear Regressions (linear, logistic etc.), SVM's, Decision trees for classification of groups and analyzing most significant variables.
- Converted raw data to processed data by merging, finding outliers, errors, trends, missing values and distributions in the data.
- Worked on different data formats such as JSON, XML and performed machine learning algorithms in Python.
- Built Analytical tools from data collected through both live streaming and batch processing using Data bricks platform.
- Experience in using various Data Visualization tools like Qlik, Tableau, Data bricks to provide Business Intelligence on simple terms.
- Implementing analytics algorithms in Python, R programming languages.
- Performed K - means clustering, Regression and Decision Trees in R.
- Worked on Na ve Bayes algorithms for Agent Fraud Detection using R.
- Performed data analysis, visualization, feature extraction, feature selection, feature engineering using Python.
- Generated detailed report after validating the graphs using Python and adjusting the variables to fit the model.
- Worked on Clustering and factor analysis for classification of data using machine learning algorithms.
- Used Power Map and Power View to represent data very effectively to explain and understand technical and non-technical users.
- Written Map Reduce code to process and parsing the data from various sources and storing parsed data into Hbase and Hive using Hbase - Hive Integration.
- Used Tensor Flow, Keras, Theano, Pandas, Numpy, Scipy, Scikit-learn, NLTK in Python for developing various machine learning algorithms such as Neural network models, Linear Regression, multivariate regression, nave Bayes, random Forests, decision trees, SVMs, K-means and KNN for data analysis.
- Responsible for developing data pipeline with AWS S3 to extract the data and store in HDFS and deploy implemented all machine learning models.
- Created SQL tables with referential integrity and developed advanced queries using stored procedures and functions using SQL server management studio.
- Worked with risk analysis, root cause analysis, cluster analysis, correlation and optimization and K-means algorithm for clustering data into groups.

Environment: Python, Jupiter, MATLAB, SSRS, SSIS, SSAS, Mongo DB, Hbase, HDFS, Hive, Pig, SAS, Power Query, Power Pivot, Power Map, Power View, SQL Server, MS Access.

Client: Teradata, CA **Jan2018 – Mar 2019**
Data Analyst/Data Scientist
Responsibilities:

- Gathered, analyzed, documented and translated application requirements into data models and Supports standardization of documentation and the adoption of standards and practices related to data and applications.

- Participated in Data Acquisition with Data Engineer team to extract historical and real-time data by using Sqoop, Pig, Flume, Hive, Map Reduce and HDFS.
- Wrote user defined functions (UDFs) in Hive to manipulate strings, dates and other data.
- Performed Data Cleaning, features scaling, features engineering using pandas and numpy packages in python.
- Applied clustering algorithms i.e. Hierarchical, K-means using Scikit and Scipy.
- Performs complex pattern recognition of automotive time series data and forecast demand through the ARMA and ARIMA models and exponential smoothening for multivariate time series data.
- Delivered and communicated research results, recommendations, opportunities to the managerial and executive teams, and implemented the techniques for priority projects.
- Designed, developed and maintained daily and monthly summary, trending and benchmark reports repository in Tableau Desktop.
- Generated complex calculated fields and parameters, toggled and global filters, dynamic sets, groups, actions, custom color palettes, statistical analysis to meet business requirements.
- Implemented visualizations and views like combo charts, stacked bar charts, Pareto charts, donut charts, geographic maps, spark lines, crosstabs etc.
- Published workbooks and extract data sources to Tableau Server, implemented row-level security and scheduled automatic extract refresh.

Environment: Machine learning (KNN, Clustering, Regressions, Random Forest, SVM, Ensemble), Linux, Python 2.x (Scikit-Learn/Scipy/Numpy/Pandas), R, Tableau (Desktop 8.x/Server 8.x), Hadoop, Map Reduce, HDFS, Hive, Pig, Hbase, Sqoop, Flume, Oracle 11g, SQL Server 2012.

Client: Algorithmia, WA **Jan 2016 – Oct 17**

Data Analyst

Responsibilities:

- Successfully Completed Junior Data Analyst Internship in Confidential.
- Built an Expense Tracker and Zonal Desk.
- Identifying inconsistencies, correcting them or escalating the problems to next level.
- Assisted in development of interface testing and implementation plans.
- Analyzing data for data quality and validation issues.
- Analyzing the websites regularly to ensure site traffic and conversion funnels are performing well.
- Collaborating with Sales and marketing teams to optimize processes that communicate insights effectively.
- Creating and maintaining automated reports using SQL.
- Understood all the Hadoop architecture and drove all the meetings
- Conducted safety check to make sure that my team is feeling safe for the retrospectives
- Aided in data profiling by examining the source data
- Extracting features from the given data set and use them to train and evaluate different classifiers that are available in the WEKA tool. Using these features, we differentiate spam messages from legitimate messages.
- Created numerous SQL queries to modify data based on data requirements and added enhancements to existing procedures.
- Implemented statistical modelling techniques in Python.
- Conducted safety check to make sure that my team is feeling safe for the retrospectives
- Aided in data profiling by examining the source data
- Performed data mappings to map the source data to the destination data
- Developed Use Case Diagrams to identify the users involved. Created Activity diagrams and Sequence diagrams to depict the process flows.

Environment: Python, Mat lab, Oracle, HTML5, Tableau, MS Excel, Server Services, Informatica Power CenterSQL, Microsoft Test Manager, Adobe Connect, MS Office Suite, LDAP, Hive, Spark, Pig, Oozie.