**Anudeep A**

**Data Scientist | Machine Learning Engineer**

Technical professional with an extensive portfolio of projects, passionate for data with global experience around the spectrum in **Data Analytics** and **Data Science** for tackling challenging problems with specialization in **Machine Learning** and **Predictive Analytics**. A motivated problem solver with an aptitude for innovation and solutions development.

**SUMMARY**

* Familiar with various ingestion techniques to bring into **R, Python, Alteryx and Azure ML** environment from different big data platforms such as a **HDFS**, Hadoop
* Hands on with various Data cleansing process like handling missing values by using feature selection such as a **replacing by mean, forward or backward fill, removing entire rows or columns or values, removing outliers, normalizing and scaling data**
* Visualized data using different visualization tools **R, Weka, Azure ML and Power BI**
* Created predictive model using **supervised, unsupervised and ensemble machine learning algorithms**
* Hands on experience in implementing classification algorithms such as **KNN, Naive Bayes, Decision Trees, Clustering, Linear and Logistic Regression**
* Familiar with predictive models using numeric and prediction algorithms like **Support vector machines and Neural Networks**, and ensemble methods like **Bagging,** B**oosting** and **Random Forest** to improve the efficiency of the predictive model
* Experience in extracting the data for creating value added datasets using **Python, R, Azure** and **SQL** to analyze the behavior; to target a specific set of customers and obtain hidden insights within the data to effectively implement project objectives
* Worked on **Text Mining** and **Sentimental analysis** for extracting the unstructured data from various social Media platforms like Facebook, Twitter and LinkedIn.
* Worked on **KAGGLE** data sets and **Microsoft Azure ML** predictive models as a part of **Data science** **Boot camp**
* Experienced the full software life cycle in **SDLC**, **Agile** and **Scrum** methodologies
* Strong conceptual, analytical, and design skills with leadership qualities. Able to work with a team or individually with excellent communication skills, and ability to meet deadlines in a fast-paced work environment

**Technical Proficiencies**

|  |  |  |
| --- | --- | --- |
| * Python, R * PL/SQL, T-SQL | * Azure Blob, Data Lake * HDFS | * MS Power BI * Azure ML |
| * C, C++, C# * Core Java | * MS SQL Server, MS Excel * ODBC | * Weka * Streaming Analytics |

**Education**

**Master of Science - Computer Science** **|** Chicago State University | **3.4** GPA

**Bachelor of Technology - Computer Science & Engineering** **|** Jawaharlal Nehru Technological University| **3.3** GPA

**Professional Experience**

**Client: Questar Assessment, Inc. - Apple Valley, MN –** Mar 2017 – Present

Data Scientist / Machine Learning Engineer

**Project:** Questar Assessments Inc. is an educational assessment provider for K-12 students. There are multiple projects within the organization related to processing of student data, loading datamart from transactional systems, creating interactive reports for districts and schools and generating data files.

* Developed predictive analytics models on Python for the utility of scoring essay program and slide of the reader scores of 500,000 Pre-K to twelfth-grade students of the schools in Tennessee State.
* Performed Data preparation on a High dimensional (Big data with large volume and variety) data sample collected from the students essay tests data using in to Python
* Data preparation Includes Data Mapping of unlined data from various formats, Identifying The missing data, Finding the correlations, scaling and removing the Junk data to further process the data for building a predictive model into Apache Spark
* Data cleaning, pre-processing, imputation, transformation, scaling, feature engineering, data aggregation, merge data frames, descriptive statistics, data visualization, score assessment mapping, reporting on Tableau dashboards
* Closely working with Document Processing and Scoring Services Expert Team to find the rule sets to build a predictive model and performed visualization for getting in depth knowledge and correlation between variables
* Processed data using R programming and developed a Predictive model to predict KPI’S (Key performance indicators) such as domain level scores within ranges and Retain Ability
* Closely working with senior Artificial Intelligence Team to create and build a Machine learning layer in the final Product
* Trained Data with Different Classification Models such as Decision Trees, Random forest, Linear & Logistic Regression, KNN models to classify quartiles & predict scores
* Based on over all Statistics, Model Performance and Run Time decided Final Model
* Achieved accuracy, precision, recall in the range of 75-80 % on average for the validated models
* The program showed improvement in the essay score of 75.7% students from fall 2017 to fall 2018
* The scores were improved by equivalent of 0.1 months of schooling for each book they read, 1.4 months if they participated last year as well and by 1.2 months for students with economic disadvantage
* Creating the Desired Models in Apache Spark Environment
* Developed Dashboard and presented to higher level management using Tableau

## **Achievements**: Presented models and results to different State Clients in a conference for Business Attraction

**Technology Stack**: Apache Spark, R, Python, Azure ML, SQL, Machine Learning Microsoft Power BI, Tableau

## **What was the spectrum for Prediction?**

In depth analysis of data, calculation of KPIs from the performance metrics data and based on the current KPIs metrics predicting the KPI for Next Admin / Next Term

## **What was the action Taken?**

Closely worked with senior data scientists for analyzing the predictive model which was built using Machine Learning Techniques purely a Classification problem

**Client: Health iPASS Inc. - Oak Brook, IL -** Jan 2015 – Mar 2017

Data Scientist / Data Analyst

**Project:** Load the Integrated Eligibility System (IES) data into Human Services Data Warehouse, allowing executive office of Health & Human Services to perform data analysis, generate reports, and monitor recipient eligibility.

* Analyzed data and performed data preparation by applying historical model on the data set in AZURE ML
* Developed predictive models based on demographic, psychographic, econometric and statistical data that deliver insights related to member enrollment
* Performed Data cleaning process applied Backward - Forward filling methods on dataset for handling missing value
* Under supervision of Sr. Data Scientist performed Data Transformation method for Rescaling and Normalizing Variables
* Developed a predictive model and validate KNN model for predict the feature label
* Performed Boosting method on predicted model to improve the efficiency of the model
* Presented Dashboards to Higher Management for more Insights using Microsoft Power BI

## **Technology Environment**: R, Python, Python libraries such as NumPy, SQL Alchemy, SQL, Microsoft Power BI, Azure ML

## **What was the prediction?**

To investigate, rationalize, and recommend optimal approaches to positively impact engagement and health outcome key performance indicators, to create a predictive model to support consumer engagement within our different product offerings based of model and past data

## **What was the Action Taken when a system was built?**

A system was successfully created on the past data, medical history, health activity of customers and runs model against the historical data and get predicted label if customers are eligible for product offer or not, on the base of that send them offer

## **What was the action Taken?**

Customers should get actual offer they like and accept it which increases the numbers of consumer engagement that leads to profit the company

**Client: Cloudizen India Pvt Ltd. -Hyderabad, India -** Sep 2013 – Nov 2014

Data Analyst

**Project:** The project was about the health care financial steam line management system. Main object of this project is to identify the recipients eligible for insurance claims to submit its Quarterly data and calculate quarter’s financial ratios.

* Analyze and mapping of data, create data dictionary by using SQL and SQL Server
* Data gathering, Analyzing data, creates reports in R
* Created ER diagram and Read ER diagram
* Created databases and tables in Microsoft SQL Server, entered dataset using flat files, csv files, and from different OLEDB sources; performed SQL querying on that database
* Used Microsoft Excel for formatting data as a table, visualization in R and analyzing data by using certain methods like Conditional Formatting, Remove Duplicates, Pivot tables, Created Charts, Sort and Filter Data Set
* Performed Statistical Analysis and Hypothesis Testing in Excel by using Data Analysis Tool
* Used SQL and SQL Server for writing simple and complex queries for finding different data elements in data set
* Used Entity Relationship Diagrams and Data mapping for better understanding of dataset, Data Modeling
* Presented a dashboard for clear view of dataset to all Stake Holders
* Created the reports and data extracts from SQL server using Excel pivot as per user requirements

**Environment**: MS SQL Server, T-SQL, R, Power Pivot, MS Excel, Regression

**KAGGLE**

**State farm Distracted Driver Competition**: Given dataset of a 100,000 driver images, each taken in a car with a driver doing something in the car (texting, eating etc.) with the goal of predicting the likelihood of what the driver is doing. After first utilizing Tensor flow, Keras, and AWS p2.x16large instance to pre-process images, a pre-trained VGG-16 model, where the last convolutional output of VGG-16 is fed back into the model, was utilized. A global average pooling layer & a fully connected layer were added and uses the activation, softmax. An SGD optimizer was incorporated with a slower learning rate and momentum of 0.9. Model Version 3 resulted in the best score, 1.26, and equated to a rank of 615/1440 (a top 40% finish) on the Kaggle Public leaderboard.

# TECHNICAL SUMMARY

|  |  |
| --- | --- |
| **Data Sources** | Azure Blob, Data Lake, HDFS, MS SQL Server, MS Excel, OBDC |
| **Programming Languages** | Python, R, T-SQL, PL/SQL, Java |
| **Data Visualization** | R, Python, Weka, Microsoft Power BI, Azure Machine Learning |
| **Data Exploration** | Azure ML, MS power BI, Azure IOT hub, HD Insight, Streaming Analytics |
| **Cloud platforms** | Azure ML |
| **Repository** | GIT HUB |
| **Machine Learning Algorithms** | Classification, KNN, Linear and Logistic Regression, Random Forest, Clustering(K- means), Neural Networks, SVM, Bayesian Algorithm, Social Media Analytics, Sentimental analysis, Market Base Analysis, Bagging, Boosting |