Sandeep Kumar Are
Mobile: +91 7989520772 **LinkedIn**
E-mail: sandeepkumar.are16@gmail.com **Personal website**

CAREER OBJECTIVE

A proactive and fast learning individual seeking an opportunity to work as a dynamic data analyst utilizing analytical & methodical skills and relevant expertise to help the company achieve business goals while sticking to vision, mission, and values.

PROFESSIONAL SUMMARY

- Total **6.10+** years' experience in IT software Development field and in that **3 +** years in **Python, Machine Learning** with large data sets of **structured data, data validation, predictive modeling and data visualization.**
- Experience on **NLP** with **Deep Learning** Models.
- **Good Experience** with **Machine Learning Supervised / Unsupervised Learning algorithms.**
- **Good Experience** with **Probability and Statistics for Machine learning.**
- Adept in statistical programming language **Python**.
- Good Experience with **NLP** NLTK**/Spacy** Libraries.
- Proficient in managing entire data science project life cycle and actively involved in all the phases including data acquisition, **data cleaning**, features scaling, **statistical modeling**.
- I have experience on **Web scraping** data with **Python-Selenium.**
- I have Experience in **Python – MongoDB**.
- Extensive experience in **Text Analytics**, generating **data visualizations** using statistical programming (**Python)**.
- Good Experience on Text Pre-Processing with **Bag of Words** (BoW**), TF-IDF**, **Word2Vec**, and **Avg word2vec.**
- Good Experience on Topic Modeling LDA and NMF Techniques.
- Good Experience in Azure Portal services.
- Involved in Language Model (BERT)Implementations.

Deep Learning Concepts:

- **NLP and Transformer based models.**

Machine Learning concepts:

- **Dimensionality Reduction:** PCA and F-Regression Models.
- **Prediction Analytics.** Simple Linear Regression**.**
- **Classification techniques:** Logistic Regression, K-Nearest-Neighbor Methods.
- **Tree-based Methods**: Regression Trees, Classification Trees**.**

- **Ensemble Methods:** Bagging, Random Forests, Boosting**.**
- **Support Vector Machine:** Maximum Marginal Classifier, Support Vector Classifier.
- **Natural Language Processing with Deep Learning.**

EDUCATIONAL QUALIFICATION

- B. Tech Electronics & Communications engineering **JNTU University, India.** 2012

ROLES & RESPONSIBILITIES

- Review the dataset and keen understanding of target output to be solved.
- Taking care of dataset with all necessary steps like missing values, checking correlation.
- Performed Data Cleaning, features scaling, features engineering, feature standardization with Count vectorization /TF-IDF vectorization NLP techniques.
- Worked with different techniques from **NLTK** / **SPACY**.
- Involved in data visualization with Matplotlib, seaborn techniques.
- Evaluated models using Cross Validation.
- Addressed over fitting by implementing of the algorithm regularization methods.
- Used Principal Component Analysis and t-SNE in feature engineering to analyze high dimensional data.
- After Exploratory data analysis, analyze which model is fit for the data.
- Analyze the results with difference measurement techniques Confusion matrix, Classification reports, Accuracy, Precision, Recall etc.
- I have Good Experience on TIME SERIES ANALYSIS.

WORK EXPERIENCE

Projects:

Working as a Software Engineer (AI, ML, NLP and Python) in Svobodha Infinity Private Limited (SAVART) from February 2020 to Till Date.

Qualitative research (Iris)

Iris uses textual, visual, and graphical data to generate insights and open threads, aspects that are critical to the investment but often elusive to spot and blindsiding analysts. For example, say a multi-national IT firm has 308 subsidiaries and is suspected to have shell companies evading taxes. The probability of spotting it is difficult and time-intensive for a group of research analysts but a matter of few milliseconds for Iris. It is obvious that the subsidiary would not be tagged as a ‘shell company’ by the firm itself. So, the approach undertaken by Iris entails capturing subsidiary data, compliance of the local jurisdictions, company supplied and third-party sourced data to pinpoint the non-compliances and then make assessments on the matter. However, scope of Iris includes not just legal jargon but around 850 topics including ethics, corporate governance, brand, moat and innovation.

Roles and responsibilities:

- I have involved scrap the data scrap the data from various sites, API’s and some more resources using Python Selenium.
- Involved in working with Python-MongoDB various tasks.
- Involved in Topic Modeling with NLP.
- Worked with team on BERT Model.
- I have experience in working with Flask-Python for creation of UI.

QUANTITATIVE ANALYSIS:

The most popular form of fundamental research is to look at numerical data. Traditional analysts look at Profit & Loss statements, Balance Sheets and Cash Flow statements to make decisions, which are often based on (biased) assumptions and estimates. This led to hedge funds building ‘innovative’ trading algorithms that analyze over millions of data points to make trading decisions. Savart has gone a step further and brought this technology for long term investments i.e., using billions of data points and not just back-test the strategies but build patterns and strategies from scratch using machine learning. While this does not sound like an enormous difference, the consequence is that the system generated strategy is far less biased and is thoroughly stress tested using not only historical data but monitored and tested in real time. Concisely, we input tons of raw data points and Quant outputs a shortlisted portfolios which is ready to be pushed through to Iris.

Roles and responsibilities:

- I have involved in what are approaches are gives the best results.
- Involved in Data cleansing, Data Cleaning, Data Analysis.
- Involved in Feature selection Models.
- Involved in Forecasting Models selections to predictions.

Worked as a Software Engineer in Netpeach Technologies Pvt Ltd from February 2015 to January 31st2020.

Project Name: <u>Uncover</u>

<u>Autism</u> Project

Description:

The goal of the project is to help society mitigate the risk of autism in new born babies by identifying
more reliable causes of autism. The project will be exploring existing research, and available information with feedback(surveys) from the patient population to define more reliable causes. The purpose of this study is to provide the healthcare industry with any findings discovered during analysis.

Tools and Technologies:

- Environments Anaconda Jupiter Notebook, NLTK tools.
- Numerical and Visualization libraries NumPy, Pandas, Matplotlib, Seaborn.
- Machine Learning libraries Scikit- learn.

Projects:

<u>Loan Predictions:</u>

Given historical data on loans given out with information on whether or not the borrower defaulted (charge-off), can we build a model that can predict whether or nor a borrower will pay back their loan? This way in the future when we get a new potential customer, we can assess whether they are likely to pay back the loan.
*This is which I have completed while doing TensorFlow 2.0 exercise from UDEMY Certifications.
Worked with: Tensor Flow 2.0 , Feed Forward Neural Network(ANN), Kera’s with TF.

https://colab.research.google.com/drive/1HbwYVE5erBLsqdwO1xqqYYRX1GqB5IFL?usp=sharing

My-Portfolio(Personal Website) -
https://sandeepkumar16nlp.github.io/sandeepmyportfolion.github.io/