SHIKHA DUBEY

✆ **+918779934451** ✉ **dubeyshikha727@gmail.com**

Linkedin **Github**

Professional Summary

Completed Master in statistical data science and big data analytics and with good learning and technical exposure in the very same domain. I am willing to work more in industry based real-time solutions with python, R, and other programming skills for faster and accurate deliverables.

Skills

Languages & Databases:-

- **Python development** - NumPy, pandas, pyspark, OpenCV, matplotlib, NLTK, scikit-learn,gensium,
- Pycharm, TensorFlow, PyTorch.
- **R-language development** - dplyr, stringr, ggplot2, etc
- **Databases** - MySQL, PostgreSQL, SPSS, Tableau.
- **Frameworks**: -Flask, Keras, Spark.
- **Technologies**: Git, AWS (S3, EC2), Heroku, Docker, Django.
- **OS** - Linux and Shell scripting
- **NLP** - Tokenization, Stemming, Lemmatisation.
- Machine learning and mathematical models.
- Understanding of required statistical analysis, visualizations, and predictions.
- Data pre-processing and post-processing-based analysis.
- Public speaking and presentation skills, client communications.

Work Exposure.

Machine Learning Internship. [May2021]

Sampatanand Tech Pvt Ltd.

Project Name:- Buktec.

- Convert PDF/images of receipts/invoices using python scripts by coding.
- Annotate images on Google Cloud Platform to define various fields like various Amounts,
- GST, Totals, etc.
- Identify data set from training, validation, and testing
- Perform training
- Monitor various training metrics and improvise training methodology/data.
- TensorFlow (Machine Learning).
- TensorFlow to achieve the same or better results as compared with Google AutoML.
- Understanding Confusion Matrix of AutoML tries to get better accuracy by hyperparameter tuning.

Data Scientist Internship. [March2021-Present]

Aditya Birla Management Corporation PrivateLimited-

Time Series Forecasting - Predict Daily Crude Oil Price.

- Bivariate analysis is performed to find the relationship between each variable in the dataset with Brent Price among Themselves.
- Second, the Bivariate Analysis with Correlation.
- Although it is Time series Model, we will also check the Causation between Independent variables and dependent variables.
- The region behind the Causality is I have to take those variables that have high correlation & high causality.
- Then I have Multivariate analysis.
- To overcome with Multivariate analysis used Techniques used PCA, used Lasso.
- One way is manually adding the columns which are highly correlated and generate it new columns.

- After I have to find out optimal lags for my dependent variable, i.e., BRENT PRICE.
- Techniques used to find the optimal lags are: -VAR Models, RFE Models, ACF and PACF plots, Cross Correlations.
- Then Features Selection, Techniques used for Feature's selection are - Sequential Forward Selection, Maximum Relevance Minimum Redundancy.

Data Analyst Internship

Jan 2021-March2021

Geomechanics, Navi Mumbai.

- Collecting and Interpreting data
- Mining data from primary and secondary sources and then organizing said data in the format that can be easily read byhumans.
- Using statistical tools to interpret the datasets, Paying particular attention to trends and patterns in the dataset.
- Creating appropriate documentation that allows stakeholders to understand the steps of the data analysis process and replicate the analysis if necessary.
- Modeling and Visualizations used tools python, Tableau.

PROJECTS

Telecom Churn Prediction **[Project Link]**

- Build predictive models to identify customers at high risk of churn using machinelearning.
- Explore the possibility of machine learning for churn prediction to retrain a competitive edge in the industry.
- **Tools:** xgboost, Grid Search, seaborn, pandas, EDA, One-Hot Encoding.

Sentiment Analysis **[Project Link]**

- Build End-to-End Machine learning pipeline for preprocessing, exploratory data analysis, modeling,deployment.
- Train over Amazon review dataset using Bert Transformer model and use f1-score for evaluation of the model.
- **Tools:** Flask, Transformers, PyTorch, HTML, CSS, JavaScript, Heroku.

Weapon Detection **[Project Link]**

- Handgun, Shotgun, and Knife detection using yolov4-tiny in videos as well as images.
- Dataset for the task is manually Annotated using LabelMe and is prepared in Yolo format with txt files containing training and testing label information. Yolov4-tiny-29 pretrained model is used for training and all training is done in the cloud platform.
- Convert Model to Tflite for android devicedeployment.
- **Tools:** Yolov4, TensorFlow, h5py, OpenCV, scikit-Image.

Recommendation System **[Project Link]**

- A part of Analytics Vidhya Hackathon
- Out of 13 challenges task was to predict the next three challenges user will opt for given the first 10 challenges the user has solved.
- Approached the problem as Text Generation Problem; where a sequence of words is used to predict the next word usingRNN.
- **Tools:** Python, scikit-learn, pandas, RNN.

EDUCATION

- **M.Sc. (Data Science & Big Data Analytics), University of Mumbai.** **July 2019–May 2021**
  Relevant Coursework: Data Structures and Algorithms, Time Series, Natural Language Processing, Machine Learning, Deep Learning, Tableau.
- **B.Sc. Mathematics, University of Mumbai** **July 2016–May 2019**

Certifications and Activities

- Practical Machine Learning with TensorFlow [NPTEL] [Co-ordinated by IIT Madras[An applied Machine Learning Course jointly offered by Google and IIT Madras which covers advanced concepts of Machine Learning and Tensorflow]
- Certification in Machine Learning [Board Infinity]
- Deep Learning [Board Infinity]
- Big Data [Board Infinity]
- Certification in Perform Sentiment Analysis with Scikitlearn/Certificate of Data Science. [Coursera]
- Member of organizing team Smt. Chandibai Himathmal Mansukhani College, Ulhasnagar-3
- CHEM –FEST-2017 Inter-Collegiate Competitions 'AZO FUN'