Nikhil Kumar

- Highly motivated and passionate Applied Data Scientist with over 4.5+ years of experience in developing, maintaining and monitoring ML versions.
- Having over 3+ years of experience in MLOps.
- *Have over 3+ years of experience in data engineering, Developing ETL / ELT data pipeline and data lake using AWS cloud services and Pyspark.*
- *Have over **7 years** of experience in python programming language.*

Rapipay Fintech Pvt Ltd, Noida — Lead *Data Platform Engineer*

MAR 2022 - PRESENT.

Project 1: Developing Data Lake architecture on AWS.

- Deploying Data Lake on **AWS cloud** which can ingest, store and process data for **AI-ML**, **Analytics** and Application use-cases.
- Automated **ELT/ETL** data pipeline with alerts and monitor capabilities**.**
- Writing **AWS Glue job** in **Pyspark** and using **crawler** for data cataloging.
- Processing **batch data** with capabilities to capture **CDC** (Change in data) using A**pache Hudi.**
- Extracting CDC from data base with help of **AWS DMS**
- Integrated analytical dashboard with data pipeline using **AWS Quick Sight.**
- Developed data lake with qualities such as scalability, usability, security, high performance and availability 27*4 with downtime less than 1 %
- Helping Bussiness and apllication to get near real time data for there insights and use-cases.
- **Technologies**: Pyspark, Apache Hudi, Python and SQL.
- **AWS Cloud Technologies** : AWS Lake Formation, Glue job, Crawler, Athena, lambda, SNS, EventBridge, QuickSight & DMS.
- Setting up MLOps ML model development environment on AWS using ECR & MLflow.

Project 2 : Financial Statement generation using NLP, Machine Learning & Deep Learning.

- Classification of data from document and extracted information from it.
- Extracting Information from document using **NER**, **Word embedding, word2vec** & **Text similarities.**
- Building Classification using TensorFlow and deploying it on ML PRV environment for Business testing.
- Developed **MLOps** pipeline to deploy and train models in batch training & to monitor environment to handle model drift on AWS.
- **Technologies:** NLP, TensorFlow, AWS S3, ECR, EC2 Instance, docker image, ALB, Route 53, Lambda, Auto Scaling and Eventbridge.

Role and Responsibility:

- Mentoring team of data engineers, data scientists and data analysts.
- Feature extension of ETL Data pipeline.
- Creating high level architecture diagram.
- End-to-end development of ML Models, Deployment & Monitoring using Python & AWS.
- Requirement gathering, brainstorming with team members
- Research and creating use-cases scoping & designing.

Greater Noida West, 201009
(+91) 9871916843
thenikhilkarn@gmail.com,

Linkedin :
https://www.linkedin.com/in/nikhil-kumar-1818511a0

Github :
https://github.com/DS-Nikhil-AI

SKILLS

Python, Scikit-learn, Pandas, Numpy,NLTK, Spacy, Kears, Tensorflow, Streamlit, Flask api, AWS.

Supervised Learning Regression,Supervised Learning Classification,Unsupervised learning, PCA, Association Rules.

CNN, RCNN, Autoencoder, GAN, YOLO Object Detection.

Statistical Analysis, Sampling, Descriptive Statistics, EDA,Inferential Statistics

CERTIFICATIONS

Innovate Data Edition by Amazon Web Services(AWS) Issued on Aug 2021

Architecturing in AWS by Amazon Web Services(AWS) Issued on Dec 2020

AWARDS

Certificate of Appreciation Issued on JUL 2021 for New Hire Management

The Standout Performer Issued on Sep 2020

PUBLISHED PROJECTS

Shriram Automall, New Delhi — *Deputy Manager, Lead Data Scientist*

JAN 2020 - FEB 2022

Project 1 : ETL Data Pipeline

- Developed Automated Data ingestion pipeline which includes data cleaning module at consumption layer.
- Writing **AWS Glue jobs in Pyspark** for data ingestion and storing it in a data parquet format at S3 and RDS.
- Cataloging metadata using **AWS crawler**.
- Developed MLOps ML development environment & QA environment for model development and business testing on AWS.

Project 2: TPX (https://thepricex.com/)

- In this project 24 ML models run at backend to predict price of pre-owned vehicles, customer segment-wise price prediction and best state prediction
- **Xgboost**, **Catboost** And **LightGbm** algorithms were developed.
- Managed Models on AWS using **MlFlow** & **Perfect Flow**.
- Deployment on AWS server using **AWS S3, EC2 Instance, docker image, ALB, Route 53, Lambda, Auto Scaling and Eventbridge**.
- Also Implemented CI/CD using **Docker, Terform, AWS S3, lambda, ECR, CLI, Git commit, Unit Tests & Integration**
- **Technologies**: Python, Pyspark,SciKit-Learn, Pandas, Numpy, Flask, Streamlit,MySQL.
- **AWS Cloud Technologies** : EC2,ECR, LBA, Rote53, Glue, Crawler, Athena, S3, Lambda, Cloudwatch & RDS.

Role and Responsibility:

- End-to-end development of ML Models using Python
- Data analysis, cleaning and deployment
- Requirement gathering, brainstorming with team members
- Research and creating use-cases
- Mentoring data analysts/scientists

Progcap —Data Scientist

June 2019 - Dec 2019

Project Credit score Model 1 and Model 2:

- Statistical model was prepared with business and financial use-case.
- Decision tree was applied for **EDA**.
- Model 1:From 150 features 25 features and Model 2: From 62 features 15 features were selected via feature engineering.
- Applied feature extraction, feature pre-processing, Validation, Metric optimization and Regularization.
- Model 1**: Random forest ML algorithm** and Model 2: **Logistics Regression ML algorithm** was used to achieve objectives.

Project Sedimentation of customer reviews:

- Sentiment labels were given to 50 thousands reviews as Highly Negative , Negative,Neutral, Little Positive and Positive.
- Data sets were clean and pre-processed using **Lemmatization** , **Stemming**, POS, Validation,and countvectorizer in **NLP**.
- **EDA** was performed by observing a high increase in key words count.
- **Naïve Bayes Multinational Classifiers ML algorithm** was used to complete EDA .
- Model was prepared using **Deep Learning LSTM** .
- Accuracy of the model was 99.97 on the data set.
- Deploying ML Model on AWS using aws beanstalk &, docker image in cloud native MLOps environment. Setting Blue Green sever for model deployment in MLOps.

Elite Tyari — *Data Scientist*

Broadband frequency conversion by DFG— *Quasi-phase matched broadband difference frequency generation in the mid-infrared region using total internal reflection in a tapered gallium arsenide (GaAs) slab.*

Technologies: MATLAB

Published in Optik - International Journal for Light and Electron Optics.

Reference no: IJLEO 55768

LANGUAGES

English, Hindi, Maithili

EDUCATION

National Institute of Technology, Agartala — *B.Tech(Electrical Engineering)*

August 2010 - May 2014

GPA: 8.06

Woodbine Modern school, Darbhanga — *12th*

June 2008 - May 2009

Percentage: 88.3%

Feb 2018 - May 2019

Project 1 : To predict credit defaulters:

- 75 features were taken from 25 features.
- Applied **Random forest algorithm EDA**.
- Applied feature extraction, feature pre-processing, Validation, Metric optimization and Regularization.
- Optimized on **XGBoost algorith**m to predict credit defaulters.
- Accuracy of the model was 93% on the data set.

Project 2: To predict the cost:

- Sixty-two different types of course cost and its sell records of around 0.1 million records sets were presented.
- Applied Random forest for **EDA** and **KNN** algorithm to predict the cost of different courses.
- Accuracy of the model was 90% on a private data set.

Project 3 : To predict sale on basis of review:

- Took around 0.1 million reviews for different courses.
- Applied **lemmatization**, **Stemming**, **POS**, Validation, and countvectorizer in **NLP** and converted into around two thousand features.
- Applied KNN to sale on the basis of review.
- Accuracy of the model was 94% on a private data set.

Maas Infosolutions Private Limited — *Python Developer*

Apr 2016 - Feb 2018

Project 1: Extraction of information & attachment from EMAIL

- Developed API to extract Subject, To ,Cc and Body of the email saving it to data base.
- Downloading all the attachments from the Email saving it into SQL table in varbinary column.
- This reduces manual effort up to 50% for saving these value into tables.

Project 2: API Development for data collection and validation

- Developed API for website to populate drop downs to collect data from the form.
- Validating data and saving it into SQL DataBase.
- Better user experience and reduced load time of the page.

Maas Infosolutions Private Limited —Intern

Apr 2015 - Apr 2016

Project 1: Data extracting from Tally:

- Developed data extraction automation tool from Tally using python and best practices in Software development.
- Taking out company name, address and preparing financial statement to a common format.

Project 2: Converting PDF into images

- Converting all pages of PDF into images using **Tesseract**.Removing noises and colour.

Project 3: PoC on API development

- Developing **REST API using Flask** in python and Consuming API using Flask.

Self Project

- Predicting Emergency Vehicle from 50K size Images using CNN, RCNN, YOLO and other DL Algorithms with data size over 50K with accuracy 88%.
- Predicting type of clothing from Fashion-MNIST data images using CNN, RCNN & Other DL algorithm with accuracy of 96 %.