Jitendra **Jangid**

Data Scientist

542 G, Orchid Island, Sector 51, Gurgaon, 122001, Haryana, India

(+91) 9933915716 | jitujangid38@gmail.com | overfitter | jitendraiitkgp01 | jitendra1996

Skills

Data Science	Linear/Logistic Regression, SVM, Naive Bayes, K-Means, Random Forest, GBM, XGBoost, LightGBM, CatBoost, CNN, RNN, LSTMs
Libraries	Scikit-learn, Numpy, Pandas, Matplotlib, Seaborn, Plotly, TensorFlow, Gensim, Pattern, NLTK, SpaCy, StanfordCoreNLP
Languages	Python, R, C, Matlab, Octave
Technologies	Amazon Sagemaker, Databricks, Jupyter Notebook/Jupyterlab, Spyder, Visual Studio, Github/BitBucket, Jira, Tableau

Work Experience

ZS Associates

Gurgaon, India

Data Science Associate Consultant — *July 2018 - Present*

- **Text Analytics Products:** Working primarily in developing a robust information extraction and retrieval system to be used by various clients
  - Study Design Optimizer: Built a fully automated & end to end Natural Language Processing (NLP) pipeline that leverages state of the art models (BERT) & transfer learning to classify text and extract explicit/implicit information using syntactic and semantic relationships from protocol documents (PDFs) to make recommendations on optimizing clinical trial design. Reduced manual effort by 50-60% for protocol designing process and accelerated clinical programs through more efficient trial designs.
  - Design Intelligence: Developed an end to end product that combines entity recognition & linking (BioBERT/BioFlair), and weak supervised learning with deep neural networks to classify unstructured biomedical text from open source data (ClinicalTrials.Gov, PubMed) to expedite the clinical program planning and trial design process. Reduced time, effort & cost for clinical trial designing (weeks to hours) and automated insights to better inform future clinical trial investment.
  - Study Design Optimizer and Design Intelligence Products featured on Cision PR Newswire
- **Predictive modeling and Explainable AI Solutions:**
  - Enrollment Rate Analysis: Analyzed the impact of study design attributes and KPIs on Enrollment Rate (Patient per Site per Month) using Generalized Linear Mixed-Effects Modeling (GLMM) to account for both fixed and random effects at protocol level
  - Enrollment Rate Prediction: Built a stack of machine learning models to predict the Enrollment Rate using various study design attributes and pre-trained biomedical embeddings

Meru Cab Company Pvt. Ltd.

Mumbai, India

Data Scientist (Internship) — *May 2017 - July 2017*

- Developed a Machine Learning model to predict driver refusal after job is awarded in order to take further action
- Implemented LightGBM method (0.94 AUC) over imbalanced data, having 90% accuracy using booking details and driver's past trend
- Designed a Restful web API using the Flask microframework in Python and deployed it using WSGI server (Apache)
- Company replaced existing model with proposed one in order to avoid refusal and driver profiling owing to its feasibility and accuracy

Awards & Achievements

2016	**Finalist,** Won Gold medal in Inter Hall Data Analytics Competition amongst 20+ teams in IIT Kharagpur	*Kharagpur, India*
2018	**Finalist,** Winner of Darkode (a coding competition) in Megalith-Civil Engineering fest, IIT Kharagpur	*Kharagpur, India*
2017	**3rd Rank,** Achieved 3rd rank out of 20+ teams in Excavate Data Analytics Competition at IIT Kharagpur	*Kharagpur, India*
2018	**4th Rank,** Secured 4th rank out of 55 students in Department of Civil Engineering, IIT Kharagpur	*Kharagpur, India*
2017	**Top 5%ile,** Finished among top 5%ile out of 5000 participants in ZS Young Data Scientist Competition	*Kharagpur, India*
2015	**Scholarship,** Merit Cum Means Scholarship Received Fee Waiver for esteemed academic excellence	*Kharagpur, India*

Competitions

Kaggle: COVID-19 Open Research Dataset Challenge (CORD-19)

Gurgaon, India

Know-Corona (Q&A System) | COVID-19 Biomedical Semantic Search — *June 2020*

- Built a transformers (BERT) based Question & Answering pipeline to answer all task questions after analyzing COVID-19 articles abstracts
- Applied TF-IDF Rank function with Pre-trained BERT Q&A Model (SQuAD 1.1) and fine-tuned on BioASQ 6b dataset (4,772 Qns) to provide insights
- Achieved 82% Accuracy & 84% F1 score of BERT BioASQ fine-tuned model on evaluation dataset (BioASQ 4b : 3,266 Questions)
- Developed an Interactive Web App (Real-Time Prediction) using Flask & React module, which answers all task questions
- Built a biomedical semantic search web application using contextual stacked embeddings of Flair and Elmo (PubMed)

Data Tales - Beyond Infinity

Chennai, India

Great Lakes Institute of Management - 3rd Rank/2000 — *Jan. 2017*

- Created a machine learning based strategy to predict the future leads and characterize the campaigns for ad-campaign data
- Applied Gradient Boosting Ensemble model (RMSE ~ 93); Presented the final model & business insights amongst 15+ teams

Ultimate Student's Hunt Machine Learning Competition

Kharagpur, India

Analytics Vidhya - 9th Rank/2500 *Oct. 2016*

- Built a Machine Learning model to predict the footfall of a park on a particular day in country Gardenia
- Achieved a RMSE score of 93.8 on private leaderboard of Analytics Vidhya using a stacking model of GBM, XGBoost and ANN

Machine Learning Challenge-2 | Funding Successful Projects

Kharagpur, India

HackerEarth - 29th Rank/6000 *June 2017*

- Analyzed the projects data (1GB) of a community and forecasted the probability of sucessfully funded project using machine learning
- Trained an ensemble model of LightGBM & CatBoost on Google-Cloud and finished with 73% accuracy on the private leaderboard

Projects

Python Package for BioBert Embeddings

Gurgaon, India

Project Link *April, 2020*

- Created a simple Python package to extract Token & Sentence level embeddings from BioBERT model (Biomedical Domain)
- Package takes care of OOVs (out of vocabulary) inherently and require only two lines of code to get token/sentence level embedding for a text sentence

Digit Writing - Hand Gesture Recognition

IIT Kharagpur, India

Soft Computing Tools in Engineering | Prof. S K Barai *March 2017 - April 2017*

- Used real time Accelarometer data from android app IMU + GPS-Stream and extracted Skewness, Kurtosis and Wavelet features
- Achieved 80-85% accuracy on validation set by applying ANN, SVM, kNN and XGBoost classification machine learning algorithms to predict digits between 0-9

Automatic Detector of Cracks in Concrete Using Deep Neural Networks

IIT Kharagpur, India

B.Tech Project | Prof. Amit Shaw *July 2017 - Dec. 2017*

- Collected 20000 images dataset of cracked concrete and classified these into 5 classes (cracked part, chalk part, joint part, surface part, others
- Trained a CNN model with 3 convolutional layers on Floydhub having 82% validation accuracy to detect cracks in concrete

Education

Indian Institute of Technology, Kharagpur

Kharagpur, West Bengal

B.Tech (Civil Engineering) *July 2014 - July 2018*

- CGPA: 8.42/10

New Public Senior Secondary School

Kota, Rajasthan

RBSE (CLASS XII) *July 2012 - July 2013*

- 82.2%

Tagore Senior Secondary School

Kuchaman City, Rajasthan

RBSE (CLASS X) *July 2010 - July 2011*

- 88%

Publications & Conferences

Improving the Volatility Forecasts of GARCH Family Models with RNNs

Mysore, India

Published in IUP Journal of Computer Sciences | 6th International Conference In Finance and Banking *Aug. 2017*

- Enhanced forecasting performance of ARCH/GARCH family models with RNN model to evaluate the volatility of daily returns of NSE
- Decreased RMSE value (TGARCH 13%, EGARCH 14%) by analyzing Recurrent Neural Network based on multilayer preceptrons

Corporate Credit Default Prediction Via Ensemble learning

Kharagpur, India

2nd International Conference on Financial Markets and Corporate Finance, VGSOM, IIT Kharagpur *July 2017*

- Developed default prediction model using ANN, XGBoost and H2O GBM algorithms and compared them with Accuracy and AUC scores
- Increased model accuracy to 92% by using conceptualized ensemble technique on 180+ distressed companies dataset

Text Classification Using Weakly Supervised — Deep Neural Networks

Gurgaon, India

Published - Medium *March 2019*

- Published an article at Medium on Text Classification Using Weakly Supervised — Deep Neural Networks
- Reference: Yu Meng, Jiaming Shen, Chao Zhang, Jiawei Han, "Weakly-Supervised Neural Text Classification", CIKM 2018

Coursework Information

Data Science Analytics Edge (Edx), Machine Learning (Coursera) , Deep Learning A-Z (Udemy), Sequential Models (Coursera)

Others Probability & Statistics (MA20104), Data Analytics (CS40003), Introduction to R & Python (DataCamp)