**Rohit Shah**

**Data scientist / Machine learning engineer**

|  |  |
| --- | --- |
| Cell: 7021081708 | **[Portfolio](https://rshah1990.github.io/)** |
| Email: shahrohit.1990@gmail.com | [**GitHub**](https://github.com/rshah1990)[**LinkedIn**](http://www.linkedin.com/in/rohit-shah-00579175) |
| Mumbai | [**Kaggle**](https://www.kaggle.com/rshah1990) [**Analytics Vidhya**](https://datahack.analyticsvidhya.com/user/profile/shah27) |

**PROFILE SUMMARY**

* Rohit is an Assistant manager in CitiusTech’s Data science team with 8 years overall experience.
* Recognized for his ability to envision and translate that vision into reality using intersection of Data science, Software engineering, and Design thinking
* Helping business with problem solving using his below skills-
  + Ideation- Converting business problem into technology problem, Roadmap creation, project planning, rapid prototyping
  + Design & development-
    - Software development (Architecture design, coding using python, SQL/PLSQL)
    - Hands-on experience in machine learning with different set of problems like Classification, Regression, Natural language processing (NLP), Time Series, Image
    - Played key role in development of clinical NLP product (Data annotation strategy, Entity extraction, semantic search, NLU) & successfully delivered multiple customer project based on Machine Learning
    - Machine learning operationalization (Drift detection, experiment tracking, CI/CD)
  + Presentation and storytelling
* 271st person in the world to clear GCP machine learning engineer certification
* Hands-on experience with all major cloud services like GCP, AWS, Azure & Databricks

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **Environments** | Windows, Linux, Cloud (GCP, AWS, Azure, Data bricks), Oracle Retail & SCM, Oracle DB, Git, GitHub |
| **Languages** | Python, SQL, PL/SQL, Core Java, Unix, PySpark |
| **Domains** | Healthcare, Supply chain management |
| **ML & DL Models** | Deep neural network (DNN), convolution neural network (CNN), recurrent neural network (RNN), BERT, Random forest, XG boost, KNN, SVC |
| **Libraries used** | * TensorFlow, Keras, pytorch, Scikit learn, Pandas, NumPy * Tensor Board, Matplotlib, Seaborn, Orange, google facets * LIME, SHAP, Eli5, what-if tool * NLTK, Genism, Spacy, sci-spacy, hugging face * mlflow, TFX, Hyperopts, fb prophet, wandb |

**Professional Experience**

* **Network Provider matching:** 6 million claims per year requires manual matching which cost around $1 million. Developed random forest model which can automate 60% of no match claim with ~1.7% of error rate which is less than human error rate of ~4% reported by business. Same model is used as recommendation system for matched claims with 90% accuracy in top 5 recommendation.
* **Clinical NLP accelerator:** The product is based on machine learning and NLP helped to abstract data locked in structure/unstructured notes (Medical records/Charts). It broadened the horizon of medical records/charts processing and at the same time minimized most of manual efforts alongside improving efficiency by fastening the data abstraction process and presenting cost effective solution
* **Medical code matching:** Client was specializing in developing, managing and licensing medical vocabulary. Target was to improve code set mapping efficiency (400%) from 15 concepts/hr to 60 concepts.hr through recommendation system. Two model was developed for problem (95% accuracy in top 5 recommendation) and procedures (92% accuracy in top 5 recommendation), Features was created using BERT and KNN model was used for recommendation.
* **Claims Denial Management:** The objective of this project was to help the client develop machine learning models, to predict claims which can get denied due to medical necessity prior to submission to payer. Dataset was highly imbalance where only 0.05% of data belongs to positive class. XG boost model was developed with 95% precision & helped client to achieve the desired success criteria for the project.
* **Machine learning model monitoring Framework:** Developed model monitoring framework to monitor evaluation metrics, business outcome, drift detection, model explanation & Infrastructure monitoring. Played role of product owner & was leading team of 4.
* **CPT & ICD code embeddings:** Almost all healthcare dataset consists of CPT & ICD codes. Developed embeddings (word2vec & Poincare) which is used in 4 different projects.

**Rapid prototyping for customers**

* **DICOM de-identification using GCP:** DICOM files consist of PHI information in metadata & burn in text. GCP provides service to de-identify DICOM files & integrate that with PAC systems.
* **FHIR integration with NLP:** Integrate FHIR with NLP which will help client to analyze unstructured information present (ICD/SNOMED code) in FHIR entity type document reference (Clinical notes).
* **Extract ontology from clinical phrases:** To extract lower level entities like laterality, anatomy, findings, diagnosis, procedure, observation from clinical phrase.

**Machine learning competition & open source contribution**

* Predicting probability of success of new grocery store (1st price)
* Capgemini Tech Challenge (3rd position)
* Predicting promotional probabilities (Top 3%)
* Tag website according to URL
* TGS salt identification challenge
* Pork price prediction
* Raised an issue with sci-kit learn for SVC model with degree three which is accepted as bug
* Reported issue for what-if tool (developed by google brain team) which was accepted as bug
* Analysis of IG Sequence to find cause of multiple sclerosis

**Course undertaken**

* GCP machine learning engineer, Year 2021
* GCP Data engineer (Coursera & Linux Academy), Year 2020
* Scalable machine learning on Big data using Apache Spark (Coursera), Year 2020
* Data camp certificated Data scientist, Year 2016
* Oracle SQL & PL/SQL developer, Year 2014
* DEV288x: Natural language processing by Microsoft (Advance NLP)

**Work History**

|  |  |  |
| --- | --- | --- |
| **Employer** | **Title** | **Dates of employment** |
| CitiusTech | Assistant manager of Data science | May 2019 – till date |
| Deloitte USI, Mumbai, India | Consultant (Oracle ERP & ML developer) | Feb 2017 - May 2019 |
| Tata Consultancy services | Oracle Apps Developer | March 2013 – February 2017 |

**EDUCATION**

Bachelor of Engineering from Mumbai University in the year 2012, India.