**Rohit Shah**

**Data scientist / Machine learning engineer**

|  |  |
| --- | --- |
| Cell: 7021081708 | [**GitHub**](https://github.com/rshah1990) |
| Email: shahrohit.1990@gmail.com | [**LinkedIn**](http://www.linkedin.com/in/rohit-shah-00579175) |

**PROFILE SUMMARY**

* Rohit is an Assistant manager in CitiusTech’s Data science team. He has total experience of around 8 years.
* 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 with NLP (Data annotation strategy, Entity extraction, semantic search, NLU) as well as structured data.
    - Machine learning operationalization (Drift detection, experiment tracking, post production support, CI/CD)
  + Presentation and storytelling
* Hands-on experience with all major cloud services like GCP, AWS, Azure & Databricks
* Played key role in development of clinical NLP product & successfully delivered multiple customer project based on Machine Learning and NLP.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **Environments** | Windows, Linux, Cloud (GCP, AWS, Azure, Data bricks) |
| **Languages** | Python, SQL, PL/SQL, Core Java, Unix |
| **Database** | SQL, MySQL, Oracle |
| **Development Tools** | Git, GitHub, Jupyter Notebook, Spyder, putty, winscp |
| **Domains** | Healthcare, Supply chain management |
| **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, scispacy, hugging face (BERT) |

**Professional Experience**

**CitiusTech Healthcare Technology Private Limited Duration: May 2019 – Till Date**

CitiusTech is a specialized provider of healthcare technology and business process services to healthcare technology companies, healthcare providers, managed care organizations, health plans, and disease management companies

|  |  |
| --- | --- |
| **Project** | **Network provider matching** |
| Summary | 6 million claims per year requires manual matching which cost around $1 million. Out of 6 million 40% of claims will never match to provider since user spend more time to find right match which they will never find.  In Phase 1 business wanted to automate no match claims, with current model we could automate 60% of no match claim with 3.2% of error rate which is less than human error rate of ~4% reported by business. |
| Responsibilities | * Building Models & communicate results with business stack holders * Collaborate with Dev-ops & application team to integrate ML model with existing system |
| **Project** | **Clinical NLP Product** |
| Summary | 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. |
| Responsibilities | * Setting up data annotation tool (Doccano) & working with BA to create annotated data * Resolving inter annotation variance & define flow to convert annotated data to BIO tagging * Creating machine learning models like RNN to extract four entities disease medication, observation & procedure. * Mapping extracted entities with concept code like SNOMED, ICD, CPT codes |
| **Project** | **Medical code matching utility** |
| Summary | 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. |
| Responsibilities | * Understood business case & conceptualizing the solution * Developed machine learning models * Shared findings with all business stockholders |
| **Project** | **Claims Denial Management** |
| Summary | Client was a leading provider of Revenue Cycle Management (“RCM”) and Physician Advisory Services (“PAS”) to healthcare providers. The objective of this project was to help the client develop machine learning models, to automatically tag claims as Approved/Denied prior to submission to payer due to medical necessity. Dataset we were dealing with 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 |
| Responsibilities | * Developed various classification machine learning models for biased distribution of target variable using Random Forest/XG Boost/Logistic Regression in Python * Explored various sampling techniques for biased distribution such as SMOTE, up sampling, down sampling and combination method * Used NLP Word2Vec for Medical coding as a feature to the model to increase the model accuracy   Achieved high precision of 95% on the claims denial model and helped client to achieve the desired success criteria for the project |
| **Project** | **ML-Ops Framework (HIMSS)** |
| Summary | The idea of this project was to develop a framework which helped the client to monitor machine learning models in production. Framework included model monitoring, model performance monitoring, drift detection and model explanation. |
| Responsibilities | * Researched about drift detection and model explanation methods * Collaborated with UX team to develop prototype * Developed presentation on proposed CitiusTech Solution * Presented to internal stakeholders to get go ahead on this project * Played role of product owner & was leading team of 4 |

**Rapid prototyping for customers**

* DICOM de-identification using GCP
* CPT & ICD codes embeddings (Reusable component)
* FHIR integration with NLP
* Automatic interpretation of echo-cardiogram
* Analysis of IG Sequence to find cause of multiple sclerosis
* Duplicate Question detection
* Predicting probability of success of new grocery store

**Machine learning competition & open source contribution**

* Predicting promotional probabilities (Top 3%)
* Tag website according to URL:
* TGS salt identification challenge:
* Capgemini Tech Challenge (3rd position):
* 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

**Course undertaken**

* GCP Data engineer (Coursera & Linux Academy), Year 2020
* Scalable machine learning on Big data using Apache Spark (Coursera), Year 2020
* Microsoft NLP certification (EDX Dev 288x), Year 2020
* ML-ops on GCP (Coursera), Year 2020
* Oracle SQL & PL/SQL developer, Year 2014
* Data camp certificated Data scientist (19 certificates), Year 2016
* Google Deep learning crash course, Year 2016

**Work History**

|  |  |  |
| --- | --- | --- |
| **Employer** | **Title** | **Dates of employment** |
| CitiusTech | Assistant manage Data science | May 2019 – till date |
| Deloitte USI, Mumbai, India | Consultant | Feb 2017 - May 2019 |
| Tata Consultancy services | Oracle apps Technical developer | March 2013 – February 2017 |

**Profile**

* **Analytics Vidhya** : <https://datahack.analyticsvidhya.com/user/profile/shah27>
* **Kaggle** : <https://www.kaggle.com/rshah1990>

**EDUCATION**

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

**PERSONAL DETAILS**

Date of Birth : 9th Nov 1990

Status : Single

Languages Known : English, Hindi, Gujrati