**Rohit Shah**

**Senior** **Data scientist**

|  |  |
| --- | --- |
| 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**

* Assistant manager at CitiusTech’s Data science team with 8+ years of experience.
* Lead projects that provided total saving of $1 million for leading claim negotiation customer.
* Pivotal role in development of clinical NLP product which extracts clinical entity from Medical notes.
* Proficient in design & development
  + **Ideation**- Adept at translating business problem into technology problem, project planning, rapid prototyping
  + **Software development** - Architecture design & hands-on experience in machine learning with different set of problems like Classification, Regression, Natural language processing (NLP), Time Series, Image
  + **Machine learning operationalization** - Drift detection, Experiment tracking, CI/CD
  + Hands-on experience in cloud services like GCP, AWS & Databricks
* Certified GCP machine learning engineer & helping 7 others to prepare.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **Environments** | Windows, Linux, Cloud (GCP, AWS, Data bricks), Oracle Retail & SCM, Oracle DB, Git, GitHub, Docker, KubeFlow |
| **Languages** | Python, SQL, PL/SQL, Core Java, Unix, PySpark |
| **Domains** | Healthcare, Supply chain management |
| **ML & DL Models** | Deep learning (DNN, CNN, RNN), BERT, Random forest, XG boost, KNN, SVC, Statistical Modelling, Probability techniques |
| **Libraries** | TensorFlow, Keras, pytorch Scikit learn, Pandas, NumPy, Tensor Board, Matplotlib, Seaborn, WEKA, Orange, google facets, LIME, SHAP, what-if tool, NLTK, Genism, Spacy, sci-spacy, hugging face, ML Flow, 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 automates 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).
  + Model processes clinical notes in less than 1 sec to extract multiple entities with attributes with F1 score of 82%.
* **Medical code matching:** 
  + Improved code set mapping efficiency (400%) from 15 concepts/hr to 60 concepts/hr through recommendation system for client specialized in medical vocabulary.
  + Two ML model (KNN) was developed for problem (95% accuracy in top 5 recommendation) and procedures (92% accuracy in top 5 recommendation), Features was created using Clinical BERT.
* **Claims Denial Management:** 
  + XGboost model was developed with 95% precision to predict claims denied due to medical necessity prior to submission to payers.
  + Dataset was highly imbalance where only 1% of data belongs to positive class.
* **Machine learning model monitoring Framework:**
  + Managed 4-member cross functional team to create model monitoring framework for HIMSS to monitor evaluation metrics, business outcome, drift detection, model explanation & Infrastructure monitoring.
* **CPT & ICD code embeddings:**
  + Created embeddings (word2vec & Poincare) for CPT/ICD codes which is used in 4 different projects.

**Rapid prototyping for customers**

* **Extract ontology from clinical phrases:** Trained NER model to extract lower level entities like Severity, body structure, Side, Substance, findings, diagnosis, procedure, observation from clinical phrase with F1 score of 92%.
* **FHIR integration with NLP:** Lead team of 6 overseeing integrate FHIR with NLP which will help client to analyze unstructured information present (ICD/SNOMED code) in FHIR entity type document reference (Clinical notes).
* **DICOM de-identification using GCP:** workflow was created for customer to de-identify DICOM files & integrate with PAC system.

**Machine learning competition & open source contribution**

* 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
* Predicting probability of success of new grocery store (1st price)
* CitiusTech Hackathon- Decentralized AI (1st price)
* Capgemini Tech Challenge (3rd position)
* Predicting promotional probabilities (Top 3%)
* For more [projects & hackathon](https://rshah1990.github.io/#projects)

**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 | Mar 2013 – Feb 2017 |

**EDUCATION**

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