KARUN THANKACHAN

kthankac@cs.cmu.edu | (412) 626-4519 | www.linkedin.com/in/karunt/

EDUCATION

Carnegie Mellon University (CMU) - School of Computer Science **Pittsburgh, PA**
Master of Computational Data Science | GPA: 4.0 *Aug 2019 - Dec 2020*
Coursework: Introduction to Deep Learning, Neural Networks for NLP, Machine Learning for Large Datasets (**Teaching Assistant**), Cloud Computing (**Research Assistant**), User-Centered Research and Evaluation, Data Science for Product Managers (**Teaching Assistant**), Interactive Data Science.

National Institute of Technology, Calicut (NITC) **Calicut, India**
Bachelor of Technology in Computer Science and Engineering | GPA: 9.3/10 *July 2011- Dec 2015*
Awarded Best Undergraduate Research Project, 2015 for an ML-based project on Sign Language Recognition. [Paper Link]

WORK EXPERIENCE

Amazon.com, ATeam **Seattle, USA**
Data Scientist 1 Intern *May 2020 - Aug 2020*
- Experimented with ETS, ARIMA, Prophet, and DeepAR for forecasting transactional volumes of Ring products to allow the business to maximize free cash flow. Achieved MAPE of 9% with a SARIMA model.
- Developed an ETL pipeline hosted on AWS Lambda, S3, and DynamoDB to automate the month-end close process for the Ring Accounting Team which created savings of 213 manual hours annually.

Dell Technologies **Bangalore, India**
Software Development Engineer 2 *July 2015 - July 2019*
- Led five developers across sales, fulfillment, and delivery teams to develop an ETL pipeline on Cloudera Hadoop and an XGBRegressor model that improved accuracy of Estimated Delivery Date (EDD) by ~37%.
- Mapped application logs to a Finite State Machine and used Bayesian networks to help identify root-cause (accuracy ~85%) of application errors reducing issue resolution time by 50% (Patent Pending).
- Implemented an ensemble model (Logistic Regression, XGBoost, and SVM) to predict (accuracy ~92%) and proactively act on issues in order processing thereby improving customer experience (Patent Approved #10713706).

ACADEMIC PROJECTS

Personalizing Education **Carnegie Mellon University | Jan 2020 - Present**
- Conducting research (**Research Assistant, Learn Labs**) under Prof. John Stamper on the effectiveness of automated question generation methods (based on BERT) for adaptive learning in introductory programming courses.
- Automated pre-requisite concept mapping using heuristics that represents the semantic, contextual, and complex nature of concepts. Pre-requisites were identified using random forests with an F1 score of 0.7. [GitHub]

Suggestion Mining **Carnegie Mellon University | Feb 2020 – Apr 2020**
- Reimplemented existing state-of-the-art for suggestion detection using parallel feature extraction layers comprising of BERT embeddings, Glove-Cove embeddings & CNN with attention to achieve an F1-score of 77.78%.
- Designed and developed a Shared Parameter LSTM with an adversarial loss to leverage multi-task learning for detecting suggestions with an F1 score of 77.72% (equivalent to the state-of-the-art performance). [GitHub]

Deep Learning Projects (CV/SR/NLP) **Carnegie Mellon University | Jan 2020 – Apr 2020**
- *Computer Vision*: Developed a face verification system using Resnet-34 (built from scratch) with an AUC score of 0.91.
- *Speech Recognition*: Identified phoneme phrases from Mel spectrogram frames using a 3-layer BiLSTM and beam search to achieve an average Levenshtein distance of 8.38.
- *Language Modelling*: Built a speech to text transcription system using a Pyramidal BiLSTM encoder, attention-based decoder, Teacher Forcing, and Locked Dropout to achieve an average Levenshtein distance of 11.4.

Cloud Computing Project **Carnegie Mellon University | Sept 2019 – Nov 2019**
- Built an ETL pipeline for ~1 TB Twitter data set using Spark and MapReduce. Engineered features leveraging tweet content and user interactions to provide suitable user recommendations.
- Designed, developed, and optimized a performant user recommendation service using AWS services (Fargate, Aurora, Elastic Load Balancer) to achieve a throughput of 4000 RPS (latency: 12ms).

SKILLS

Languages/Libraries	Python (pandas, NLTK, Scikit-learn, PyTorch, TensorFlow), Java, C, SQL, JavaScript
Frameworks/Technologies	SageMaker, AWS, GCP, Spark, Hadoop MapReduce, Hive, Impala, Tableau, Git