Divyansh Madhwal

Email : divyansh.madhwal@gmail.com
Mobile : +91-962-565-5186
www.linkedin.com/in/divyansh-madhwal
www.github.com/Divnsh

EXPERIENCE

- **Easyrewardz Software Services Pvt. Ltd.** — Gurgaon, HR
  Data Scientist — *May. 2018 – Feb. 2020*
    - **Insights Creation Engine**: Created an engine for generating customer level customizable messages with key indicators calculated in the backend, used by store managers for better customer service. Indicators include Customer Lifetime Value (BGNBD-GammaGamma model), product recommendations (Alternating Least Squares), churn prediction (Gradient Boosted Trees), sentiment analysis (Passive Aggressive Classifier), and several non-machine-learning based insights. *Languages and technologies*: PySpark, Python, Spark, SQL, Cassandra, Elasticsearch.
    - **Tags Creation Engine**: Modified old and created new modules. The engine selects customers based on customizable inputs selected by the user. *Languages and technologies*: PySpark, R.
    - **Segmentation**: Designed and developed an analytics application using Flask API and python. Segments and ranks customers, stores and products based on the features and their priorities selected by the user (K-means and agglomorative hierarchical clustering).
    - **Single View Creator**: Created a customizable customer level single view creator in PySpark for data science applications, significantly increasing the productivity of the team by automating feature engineering on big data.
    - **Forecasting KPI's**: Performed time series forecasting on KPIs such as weekly sales, number of newcomers, total discounts availed, etc. (Holt-Winters, Hierarchical Prophet, LSTM).
    - **Fraud Engine**: Designed and developed a fraud detection engine that detects anomalies at a range of specified window sizes (weekly, biweekly, monthly, etc.) (Local Outlier Factor, DBScan, Isolation forest, Variational Autoencoder). Extracted rules from predictions using decision tree algorithm. *Languages and technologies*: PySpark, Python, Spark, SQL, Cassandra, PyTorch.
    - **Lapsation**: Implemented BGNBD model for predicting number of days till next purchase for each customer. Used in identifying lapsed customers.
    - **Propensity Modelling**: Created a model for generating propensity scores for customers indicating their propensity of arrival in campaign periods(Gradient Boosted Trees). Optimized threshold based on F1-scores. Cumulative accuracy profile used to assess the discriminative power of the model. Improved hit rates from 5% to 25%. Co-created a second model for predicting the discount levels to be availed by each customer (Feed forward neural network). *Languages and technologies*: PySpark, Python, Spark, SQL, Tensorflow.
    - **Customer Delight Index**: Prototyped a scoring mechanism for calculating an index representing customer delight through answers in feedback forms.
    - **Hit-rate Confidence Intervals**: Used Gaussian Mixture Models, ANOVA and bootstrapping techniques for calculating confidence intervals for click rates on offers during the 6 time slots in a day, considering offer type and mode of communication.
    - **Brand Analytics**: Performed descriptive analytics for a major brand at various levels of aggregation.

- **National Institute Of Public Cooperation And Child Development** — Delhi
  Intern — *May. 2013 – Jul. 2013*
    - **Vital Statistics**: Performed analysis, interpolation and extrapolation of data related to vital statistics and education in various states of India.
    - **Management**: Assisted in management of an NGO leaders training workshop, hosted for NGO's from diverse locations in India.

EDUCATION

- **Manipal Academy Of Higher Education** — Bangalore
  Post Graduate Diploma in Data Science (Full Time); CGPA: 8.82/10 — *Jul. 2017 – Jun. 2018*
    - Coursework included R Programming, Statistics, Big Data, Machine Learning, Data Visualization, Unstructured Data Analysis, Databases, and Finance.

- **Institute And Faculty Of Actuaries** — London
  CT-5:Contingencies; CT-2:Finance and Financial Reporting; CT-6:Statistical Methods; — *Jun. 2016 – Dec. 2017*
  CT-8:Financial Economics

- **Institute of Actuaries of India** — Mumbai
  CT-3:Probability and Mathematical Statistics; CT-1 Financial Mathematics; — *Oct. 2014 – Oct. 2015*
  CT-7:Business Economics

- **P.G.D.A.V. College, University Of Delhi** — Delhi
  Bachelor of Science (Hons.) in Statistics — *Jul. 2011 – Jan. 2015*

- **Online Courses**

  Introduction to Deep Learning; Bayesian Methods in Machine Learning; Natural Language Processing; Deep Learning in Computer Vision : **Higher School of Economics, Coursera**, *2020*
  Deep Learning Specialization: **deeplearning.ai, Coursera**, *2019*
  Regression Modeling in Practice; Machine Learning for Data Analysis; Data Management and Visualization; Data Analysis Tools : **Wesleyan University, Coursera**, *2018*
  M001 MongoDB Basics : **MongoDB University**, *2018*

Projects

- **Optical Character Recognition**: The app takes one or more images as input and returns docx files containing text detected in the images. Supports English, Tamil, Hindi, Telugu, Bengali and Malayalam. OpenCV used to pre-process the images. Pretrained FastSRCNN models used to upscale small images. Legacy and/or LSTM Tesseract models are used based on availability for the language. Web app built using Dash. *https://img2doc.herokuapp.com/.*

- **CTR prediction**: Data analysis and click-through-rate prediction on a travel agency website data using classification algorithms like Logistic regression, LightGBM, etc.

- **Recommendation Systems**: 1). Recommend challenges on a competition hosting website using multi-VAE (uses multinomial likelihood) with feature embeddings of challenges (extracted from encoder of another VAE). Mean Average Precision used as performance metric. Beta, weighting the KL Divergence loss, tuned to ensure the MAP values do not collapse after a few epochs. *Based on paper: https://arxiv.org/abs/1802.05814.*
  2). Joke ratings are predicted per user using Hybrid-VAE. Mean word embeddings for jokes are extracted from Bert-base model and reduced from 768 to 3 dimensions using VAE. PCA with 3 components is performed on other joke features and merged with reduced embeddings. Embedding layer in the H-VAE contains the above embeddings, and is used in training the model. H-VAE provides an RMSE of 4.01 compared to SVD's 4.29 on validation set. *Based on paper: https://arxiv.org/abs/1808.01006.*

- **Chatbot**: Conversational chatbot to query StackOverflow threads. Starspace embeddings (trained on duplicate detection task) used to encode user input, StackOverflow threads. User-intent (2 types: dialogue/query) and tag classifiers trained with logistic regression and one-vs-rest logistic respectively. Pre-trained neural net engine - Chatterbot - carries the conversation in case user-intent is dialogue. Question is classified as a tag (a programming language), the thread embedding under that tag - closest to query embedding - is the basis for response. Telegram API used to deploy the bot. *https://rb.gy/ivkhrg.*

- **Depression Detection Dashboard**: Plain text or a single-column csv/excel file is taken as input for the model which predicts the depression score.Top 3 latent depression topics are extracted from the plain text. For training, text is scraped from reddit via PRAW with multithreading. Text pre-processing is performed with multiprocessing. RoBERTa's 97% accuracy beats XGBoost's 85% on TF-IDF and LSA features. Gensim's LDA mallet model maximizes coherence score to find relevant topics in depression posts. The containerized web application is built in Dash and deployed with gunicorn and nginx. *https://gitlab.com/divyansh.madhwal/dep-dash-app.*

Accomplishments

- **Professional Awards**: Awarded Star Performer in 2019 at Easyrewardz Software Services Pvt Ltd.

- **Entrance exam ranks - top 10%**: 68th in BHU Statistics entrance exam 2014, 153rd in IIT-JAM Statistics 2015.

Skills

- **Languages**: Python, PySpark, SQL, R.

- **Technologies**: Spark, Hadoop, Tensorflow, Pytorch, NoSQL Databases, Flask, Dash, Docker, Tensorflow-serving, Git, Ubuntu, MS Office, VBA Excel, LaTeX, Shell.

- **Analytics**: Statistics, ML, NLP, Deep Learning, Computer Vision.