**Kunal Khadke**

201-685-8153| Srikanth.p@emonics.com| San Antonio,Texas [**https://github.com/kunalkrk1**](https://github.com/kunalkrk1)

**PROFESSIONAL SUMMARY**

* Over 9+ years of experience in Predictive Modeling, AI development, Statistics, Machine Learning & Data Science
* Strong hands-on experience in Data Modeling, Data Wrangling, Statistical Modeling, Data Mining, Machine Learning, Data Visualization, and transformations using DataFrames
* Expert in python libraries such as NumPy, and SciPy for mathematical calculations; Pandas for data preprocessing/wrangling; Matplotlib, Seaborn for data visualization; Scikitlearn for machine learning, Tensorflow and Keras for Deep Learning

**SKILLS**

|  |  |
| --- | --- |
| **Programming Languages** | Python(Numpy, Pandas, Scikit-learn, Matplotlib, Seaborn Keras, Tensorflow, NLTK, Regex,OpenCV, Pytorch) SQL, Matlab, R |
| **Big Data Technologies & Tools** | Tableau, Qlik, Excel, Jupyter Notebook, RStudio, Google Cloud Platform (GCP), Git, Jenkins, Kubernetics, Docker, Linux |
| **Machine Learning** | Predictive modeling, Linear and logistic regression, SVM`s, Clustering- K-nearest neighbors, Segmentation methodologies, Decision Trees, Random Forest, Support Vector Machines, PCA, Decision Trees, Random Forests, ANN, CNN, Natural Language Processing, Deep Learning |
| **Data Techniques** | Data wrangling, ETL, Database management, Data Pipeline,  BigQuery, DataFlow, Dashboard creation and Visualization, A/B Testing, Experimentation, Statistical Analysis |
| **Databases** | Postgress, MySQL |
| **Cloud Platforms** | AWS , GCP, Azure Databricks, Apache Airflow, Kubeflow |

**PROFESSIONAL EXPERIENCE**

**AI/ML Engineer** **Nov 2023 - Present**

Victory Capital (Client) – Emonics LLC |San Antonio, Texas

**Machine learning model for financial advisor prediction**

* Developed a predictive machine learning model aimed at forecasting the purchasing behavior of financial advisors regarding the company's mutual funds in forthcoming months, while also recommending specific products. This initiative significantly contributed to augmenting sales representatives' performance, elevating sales by an impressive 20% annually
* Built and maintained an end-to-end ML pipeline using AWS SageMaker, incorporating data ingestion, cleaning, preprocessing, feature engineering, model training, and deployment
* Leveraged AWS S3 and Redshift as the primary data storage solutions, ensuring secure and efficient handling of the datasets utilized in the model development process.
* Experimented with logistic regression, random forest and XGBoost algorithm for this classification model. Deployed the XGBoost model which gave the highest AUC score
* Monitored the model drift every month to check the health of the model

**Machine learning model for customer segmentation using Clustering**

* Customer segmentation helped the company tailor investment options, strategies, advice and communication along with targeted approach leading to higher conversion rates
* Automated data ingestion using AWS Glue from different data sources and saved the data in both AWS S3 and Redshift
* Conducted thorough Exploratory Data Analysis (EDA) utilizing Matplotlib and Seaborn libraries, leveraging insightful visualizations for data cleaning and innovative feature engineering techniques.
* Implemented model deployment by utilizing the DBSCAN clustering algorithm, effectively segmenting data based on inherent patterns and similarities.
* Employed the Elbow method as a strategic tool to ascertain the ideal number of clusters, optimizing the model's performance and accuracy.
* Aligned the segmentation strategy with the business objectives, establishing a clear and meaningful direction based on the determined optimum number of clusters. This approach facilitated targeted decision-making and actionable insights derived from the segmentation analysis.

**Data Scientist** **July 2019 – Nov 2019**

Toyota Motors (Client) – Hindujatech LLC |Ann Arbor, MI

**Predictive maintenance of manufacturing tool using Machine Learning and AI**

* Created a binary classification model from the part/tool data which significantly helped the manufacturing engineers alert the tool failure in advance and reduce the tool scrap by 70% reducing unplanned downtime by 60% and increasing overall equipment effectiveness (OEE)
* Built and maintained an end-to-end ML pipeline using AWS SageMaker, incorporating data ingestion, cleaning, preprocessing, feature engineering, model training, and deployment
* Created a Docker Image of code to use it from the local machine and effectively upload it to AWS
* Created the data streamlining pipeline using Kubernetics to trigger the data ingestion time
* Leveraged AWS S3 for data storage, providing a scalable and secure solution for the vast amounts of data generated by the manufacturing equipment
* Automated data preprocessing and feature engineering steps using AWS Glue(ETL), reducing the overall time required for data preparation
* Used SMOTE method for oversampling the failure class from 3% to 50% for imbalance dataset in training
* Reduced overfitting by K fold cross validation and optimized the model parameters for reducing the false positive error in the confusion matrix
* Trained different classifications models like Logistic Regression, SVM and Decision Tree. Deployed the Decision Tree model with a Recall of 80%
* Deployed the model through an API using AWS API Gateway and Lambda, providing real-time health updates on the manufacturing tool to the maintenance team
* Conducted A/B testing to compare the performance of the predictive maintenance model against the traditional maintenance procedures, validating the effectiveness of the new approach

**Regression analysis for predicting injury**

* Implemented a Random Forest regression model to predict injuries, reduced engineering analysis time by 40%, streamlined through automation in Apache Airflow and Kubernetics
* Automated data pre-processing tasks with Kubeflow, including imputing missing values, feature engineering and modeling
* Used Linux commands to interact with AWS and other tools
* Conducted EDA and visualization using matplotlib and seaborn, identifying highly correlated factors such as barrier position and occupant position with injuries
* Using Random Forest multivariate regression predictive model with r-squared value of 0.73 helped the company avoid running crash simulations for minor feature variations, thereby reducing injury analysis time by 30%

**Vehicle Type Image Classification**

* Productionalized a vehicle image classification of Sedans, SUV`s and Pickup- Trucks using deep learning model with 85.3% Precision, reducing manual post-processing times by 20%
* Performed data augmentation techniques using OpenCV2 like rotating, flipping and scaling
* Experimented with different with CNN, Resnet51, and VGG-16 models
* Resnet51 gave the lowest validation loss and was chosen for classification
* Used Confusion matrix, AUC/ROC, Precision and Recall to evaluate model performance for the classes
* Deployed Resnet51 model using Widget and Streamlit as a user-friendly API and is in discussion for product deployment

**Data Analyst**  **Aug 2014-July 2019**

General Motors (Client) – Altair |Warren, MI

* Carried out Predictive analytics based on the simulations data to generate insights for future efficacy
* Created dashboards using Tableau to visualize patterns for assisting in vehicle design predictions
* Automate post-processing analysis activities with Python scripting
* Identified potential opportunities to build machine learning models and improve overall prediction time for injury

**DATA SCIENCE AND MACHINE LEARNING PROJECTS** [**https://github.com/kunalkrk1**](https://github.com/kunalkrk1)

**Clustering on E-Commerce Data**

* Performed data pre-processing and transformed description columns using regex and spacy to remove stopwords, punctuations and HTML tags from the data
* Tokenized the description columns using nltk library to a state that can be interpreted by the algorithm
* Developed clustering model using KMeans algorithm to get clusters of product from the unlabeled data
* Developed clustering model using KMeans algorithm to get segments of customers from the product clustered data
* Analyzed each cluster to characterize the clusters based on the cluster statistics
* Developed an ensemble-based model for classifying customers into the segments extracted through clustering

**Credit Card Fraud Detection**

* Dataset contains numerical input variables of transactions made by credit cards
* Data exploration and analysis done to help understand the data
* Isolation Forest model (accuracy score: 99.7%) and OneClassSVM (accuracy score:98.2%) used to detect anomalies in the dataset

**Gas price forecasting**

* Dataset contains gas prices per week of 5 years
* Experimented with ARIMA, LSTM and AutoKeras time series models for forecasting results

**University Management System**

* Extract required information in the university database by writing queries in SQL
* Dataset consists of student ID, teacher ID, department ID, course\_name and so on

**EDUCATION**

**Master of Science, Data Science**  2021

Eastern University| PA

**Master of Science, Mechanical Engineering** 2013

Purdue School of Engineering | Indianapolis, IN

**Bachelor of Engineering, Mechanical Engineering** 2010

Pune University | Pune, India