Hrushikesh Sahu

Data Science Trainee at AlmaBetter | Sambalpur(1 Year Experience)

hrushikeshsahu19@gmail.com | 7848866575 | linkedin.com/in/hk-sahu | github.com/hrushikeshsahu19

WORK EXPERIENCE

Automated-Q-A-System
Freelancing

03/2021 - 05/2021

Tags: Theme Extraction, NLP, spaCy, LDA, Transformer, Model Deployment

- Developed a **web application** that returns the **contextual answer to questions** related to customer inquiries skimming through a corpus of **15K+ documents**
- Implemented the **LDA topic model** algorithm on the content of the corpus to tag major topics and evaluated the **optimal topic model** using **perplexity score** and check the stability of topics across time.
- Performed general **text preprocessing** in **spaCy** such as named **entity recognition**, **lemmatization, tokenization, and vectorization using TF-IDF vectorizer**.
- Deployed TD-IDF and transformers model using **Flask API** on **AWS** which returns top 5 most relevant contextual answers from the **text corpus** with an overall relevance **score of 75%.**

PROJECTS

Meru taxi trip time prediction
AlmaBetter Verified Project

12/2020 - 01/2021

Tags: Regression, XGBoost, Gradient boosting machine, MSE, R-square, Decision tree, VIF, homoscedasticity, multicollinearity, Gridsearch CV, feature engineering, Lasso, Ridge, Pearson correlation

- Built a **regression model** using **GBM, Decision tree regressor**, and **XGBoost** models to predict **taxi trip time** in Delhi for a time period of six months
- Used a **folium graph** to visualize the pick-up and drop-off locations and used **heatmaps**, which were essential for **EDA**
- Applied **feature engineering** to obtain new features such as distance, speed, peak hours, busiest days and used **Pearson correlation, VIF values** to avoid **multicollinearity** in **Linear Regression.**
- Applied **Lasso and Ridge** regularisation for optimizing the fit of the model and used **Gridsearch CV** for **hyperparameter tuning**, which resulted in **R- square score** of **0.71** on the test dataset.

Marketing Campaign Effectiveness Prediction
AlmaBetter Verified Project

01/2021 - 02/2021

Tags : Binary Clasification, SMOTE, SHAP Interpretability, XGBoost, Multivariate Outlier Treatment, Time Series Analysis, Class Imbalance

- Developed a **stacked** model using logistic regression as a **meta-classifier** on base classifiers such as **XGBoost**, **SVM** and **RF** to predict whether a customer will subscribe to an FD scheme as a result of a marketing campaign.
- Treated **multivariate outliers** using **Isolation Forest** and applied **SMOTE** boosting on normalized data to handle class imbalance and obtained 92% **AUC-ROC** on test data.
- Leveraged **SHAP** plots to determine the most important features contributing to purchase such as number of outgoing calls, bank balance, personal loan, housing loan etc. and **increased** the customer acquisition rate by **15%.**

Loan Default Prediction
AlmaBetter Verified Project

02/2021 - 03/2021

Tags : Classification, KS Statistic, Gains table, SMOTE, Hyperparameter Tuning, SHAP Interpretability, Extreme, Gradient Boosting

- Developed an **XGBoost** binary classifier to predict whether a customer will default on a loan and achieved the **AUPRC** scores of 92% and 88% on test data respectively.
- Engineered a new class of attributes known as **decayed field** variables and developed out-of-pattern variables on historical loan and bureau data to identify risky customers and **reduced bad rate** from **14.2%** to **10.5%** .
- Performed missing value imputation using **KNN-Imputer** , implemented **SMOTE** boosting to oversample the minority class observations and carried out hyperparameter tuning using **Bayesian optimisation** .

TECH STACK

Languages
Python,Java, C, PHP, SQL, MySQL

Frameworks
Scikit-learn, Keras, Pandas, Numpy, Seasons, Matplotlib, spaCy, Keras, Pandas, OpenCV, NLTK, Plotly, Flask

Platforms
Jupyter Notebook, Google Colab, MongoDB, PostgreSQL, Spyder, Terminal, PyCharm, Oracle

RELEVANT COURSEWORK

Machine Learning by Andrew Ng (2021)
Linear/Logistic Regression CART, SVM, Naïve-Bayes, KNN, Boosting, Gradient descent, Neural networks

Machine Learning (AppliedAI) (2020)
KNN, SVM, Bagging, Random Forest, Naive Bayes, Boosting, GBDT, Xgboost, K-Means, PCA, LDA, NLP

Statistics for Data Science (Udemy) (2019)
Probability distribution, confidence interval, Hypothesis Testing, central limit theorem,Co-relation , Regression

EDUCATION

B.Tech In Computer Science and Engineering
Indira Gandhi Institute Of Technology, Sarang

2021

CGPA - 8.55

XII-Higher Secondary
Yuvodya Junior College, Balangir

2016

Percentage - 77.16%

X - Secondary
Govt. High School, Kuhibahal

2014

Percentage - 75.33%

ACHIEVEMENTS

Awarded Scholarship for Higher Education by State Government, 2020

1st in District Level Cricket Competition, 2019

INTERESTS

Competitve Coding | Blogging | cycling | Cricket