**SHUBHAM SAHU**
**M.Tech (Industrial & Management Engineering)**
shsahu@iitk.ac.in | +91 8400191929
https://www.linkedin.com/in/shubham-sahu-6ab3a2b9

## ACADEMIC DETAILS

| YEAR | QUALIFICATION | EDUCATIONAL INSTITUTE | PERCENTAGE |
|---|---|---|---|
| 2019-21 | M.Tech (Industrial & Management Engineering) | Indian Institute of Technology, Kanpur | 7.7* (CPI) |
| 2013-17 | B.Tech (Mechanical Engineering) | National Institute of Technology, Uttarakhand | 8.33/10 |
| 2013 | Class XII (U.P. BOARD) | PT. R.P.M. Inter College | 92.6% |
| 2011 | Class X (U.P. BOARD) | Saraswati Vidya Mandir Inter College | 80% |

**upto 2nd semester*

## SUMMER INTERNSHIP

**Data Science Intern at Mphasis Pvt. Ltd., Bangalore** *(May'20-Jun'20)*

- **Project 1 : Information Extraction of fields (Address, Commencement Date, Currency, Clause-Alteration)** from unstructured text documents that showed the agreements between the organizations using NLP
- Data cleaning: Extracted chunks for each page removing all the noise, stopwords using Lambda function
- **Sentence segmentation, word tokenization** and **Lemmatization** techniques were done along with **data tagging** as 0 or 1, performed **Exploratory Data Analysis (EDA)** and **Feature Engineering (FE)**
- **Imbalanced dataset** is observed which is handled using **oversampling** technique and **TF-IDF** feature extraction methodology is used
- Applied **Logistic Regression, Support Vector classifier** algorithms and compared these models on common metric
- Support Vector Classifier performed better with good recall, precision, f1-score values for each field
- Regular expressions were used to extract all the fields after classifier algorithm

**Project 2 :** *(May'20-Jun'20)*

- Objective: **Data Augmentation** for NLP using **Covid-19** Dataset
- Used **text augmentation technique (OCR, Keyboard distance)** to increase the training data on character level.
- Developed an algorithm that works on any sentiment analysis dataset with structured data with a word limit of not less than 5 to 1,00,000 words

## ACADEMIC PROJECTS

**Data Mining**

**Movie Review Sentiment Analysis** *(Oct'19-Nov'19)*

- Objective: To **predict the sentiment (Negative, Somewhat Negative, Neutral, Somewhat Positive, Positive)** of Rotten Tomatoes movie based having 1.5 lakh reviews and 4 attributes (Phrase ID, Sentence ID, Phrase and Sentiment)
- Performed data-cleaning and pre-processing including **Exploratory Data Analysis (EDA), Feature Engineering, Data Visualization** including **word cloud** for each sentiment
- Feature Extraction techniques- **CountVectorizer, TF-IDF (Term Frequency- Inverse Document Frequency)**
- Generated **classification report** & **confusion matrix** using **Logistic Regression**, **Stochastic Gradient Descent**, **Random Forest**
- **Random Forest with TF-IDF** was observed as best model with **accuracy** of **0.63**

**Applied Machine Learning**

**Prediction Modelling of bank loans using Logistic Regression using a bank Dataset** *(Mar'20-Apr,20)*

- Objective: To classify whether a loan applicant will be a defaulter at a later stage or not based on factors such as credit amount, employment, property, credit_cards etc.
- **Class 'credit_rating'** was **unbalanced** with **70%** of dataset belonging to **'good'** class
- It was observed that offers for Car loan can pick up more customers for loan from the bank
- Logistic Regression Models were used to classify the credit rating class. Final Model gave an **accuracy** of about **76%** and a **precision** of **74%** and **recall** of **63%, AUC of ROC** curve was **0.87** which shows that model predictive power is good

**Statistical Modelling for Business Analytics**

**Predicting Prices of a Real Estate using Statistical Regression Model** *(Jan'20-Feb'20)*

- Objective: To study the various **factors affecting the price of Real Estate per square feet**
- Calculated **correlation matrix,** Performed **Exploratory data analysis, Heteroskedasticity check** with **white test,** checked for **Multi-Collinearity test** using **Variance Inflation Factor (VIF)**
- Finalized a **multivariate Non-Linear regression model** on the basis of **Adjusted R square value(0.67),** Residual Plots
- Significant variables were distance from metro station, number of near convenience stores and transaction date

**Predicting Income class using Logistic Regression using Adult data set** *(Mar'20-Apr'20)*

- Objective: To predict whether a person's income is <50K or >=50K based on factors such as age, education, marital status etc.
- Data cleaning: Reduced the total no of factors in some columns and handled missing values and discrepancies
- **Logit** and **Probit** models were used for classifying the income class
- The performance was similar to an **accuracy** of about 82.3% , **precision** of 62.17% and a **recall** of 52.8%
- **AUC of ROC curve** was 0.88

**Marketing Research**

**Analysis of consumer satisfaction of JIO SIM card with introduction of IUC** *(Feb'20-Mar'20)*

- Data was collected with the help of **questionnaire**, conducted the online survey & did Analysis in SPSS
- **Research questions-** Are current Jio users going to switch to some other networks in case Jio continues to charge IUC, How much are JIO customers satisfied with Recharge Plan, Network Service, Customer Care Service and Internet Speed
- Performed **Exploratory data Analysis** on surveyed data and analysed it using the statistical test in **SPSS** to get useful insights about how much are customers satisfied with Jio services and decision to charge IUC

## COURSEWORK AND SKILLS

| | |
|---|---|
| **Relevant Courses** | **Data Mining and Knowledge Discovery** \| **Probability & Statistics** \| **Statistical Modelling for Business Analytics** \|**Advanced Statistical Methods for Business Analytics (ongoing)** \| Applied Machine Learning \| **Marketing Research** \| Introduction to Computing (JAVA) \| Operations Research \| |
| **Skills** | **Python** (NumPy, Pandas, Scikit-Learn, Seaborn, Matplotlib, NLTK ) \| **R** \|MS Office (Excel, Word, PowerPoint) \| SQL \| JAVA \| |

## POSITION OF RESPONSEBILITY

- Orientation Team Member (OTM) at Counselling Service, IIT KANPUR *(Dec'19-Jan'20)*
- **General Secretary** of Mechanical Engineering Department, NIT UTTARAKHAND *(Apr'15-Mar'16)*

## ACHIEVEMENTS AND CERTIFICATIONS

- Secured **1st** prize in cricket in **Institute Gathering-SPARKS 2015**, NIT UTTARAKHAND
- Certifications: **R** Programming A-Z (Udemy) \|Introduction to Machine Learning (Coursera) \|**Python** A-Z (Udemy)