SUNIL SINGH KUSHWAHA kushwaha@iitk.ac.in | +919918989593
M.Tech (Industrial & Management Engineering) https://www.linkedin.com/in/sunil-singh-kushwaha-66b431120/

Year	Qualification	Educational Institution	Percentage
2019-21	M.Tech (Industrial & Management Engineering)	Indian Institute of Technology, Kanpur	8.61* (CPI)
2014-18	B.Tech (Mechanical Engineering)	Harcourt Butler Technical University, Kanpur	75.92 %
2013	Class XII (CBSE)	Toolika Public school, Ghazipur, U.P.	83 %
2011	Class X (ICSE)	St. John's School, Ghazipur, U.P.	88 %

*upto2nd semester

INTERNSHIP

Data Science Intern at **Mphasis Next Lab**, Bengaluru *(May-June'20)*

A) **Objective:** A System for Entity driven Coherent **Abstractive summarization** framework that leverages entity information to generate informative and coherent summary.

- Data exploration: File download from website and extraction of text from pdf using **Apache Tika.**
- Data manipulation: Abstract, Content and Keywords were extracted and Special Characters, emails, Page Number and other unwanted were removed.
- Used three approaches like **generating summary**, **candidate sentence selection** using entity and **comparing similarity vector** to get candidate sentences and thus comparing N-grams.
- Rouge Score was achieved around 0.42
- Package used : Beautiful soup, Apache tika, NLTK, **Gensim**, Rouge, Spacy, Scipy, **GensimDoc2vec,** Regular Expressions.

B) Data Augmentation Techniques for Boosting text classification

Objective: A strategy that enables significantly increase the diversity of data available for training models, without actually collecting new data

- It is used for Image, Audio, Text(character, word, sentence level) Augmentation to increase the size of the dataset and introduce variability in the data.
- Word augmentation: Three methods were used **WordNet** (lexical database in English), **Round Trip translation** (RTT) and **Synonym** replacement**.**
- WordNet Model did not perform well on new words in English language like **COVID'19**, CORONA.

ACADEMIC PROJECTS

Data Mining

Predicting Box office revenue of a movie using Random Forest *(Aug-Nov'19)*

- This Project predicted how much revenue a movie is going to make at the box office.
- Steps include **data pre-processing**, **exploratory data analysis**, **Linear Regression, Random Forest models** building and finally predicting test data from finalized random forest model.
- For movie revenue top five important variables came out be popularity, budget-year-ratio, weekday-release, month of release, and main-genre.
- Packages: plotly, ggthemes, dplyr, stringr, ggplot2, knitr, viridis, VIM, lubridate, RandomForest.

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(0.67),** Residual Plots.
- Statistically significance variables were distance from metro station, number of near convenience stores and transaction date

Predicting Income class using Logistic Regression using Adult data set *(March-April'20)*

- Objective: To predict whether a person's income is <50K or >=50K based on factors such as age, education, marital status, gender 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 84.3% , **precision** of 61.9% and a **recall** of 52.8%
- AUC of ROC curve was 0.88

Applied Machine Learning

Movie review sentiment analysis *(Mar'20-Apr,20)*

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.**

COURSE WORK AND SKILLS

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

ONLINE LEARNING & CERTIFICATIONS

- **R Programming** A-Z: R for Data Science with Real Exercises (Udemy)
- **Machine Learning for all** by University of London offered at Coursera
- **Introduction to machine learning** by Duke University offered at Coursera.

POSITION OF RESPONSIBILITY

- **Manager**, Takeoff event at Techkriti 20, IIT Kanpur
- Member, Transportation Team at **Open House** 20, IIT Kanpur

AWARDS AND ACHIEVEMENTS

- Secured **1138** rank in GATE 2019 (Mechanical)
- Winner of **Robowars** at Mecharnival'16 (Techno-cultural Fest) at **HBTI** Kanpur.