Data Analysis.

For the purpose of testing and validation resumes were collected from 2 different sources on Kaggle. Both the data sets combined consists of 2317 data points with a combination of blue- and white-collar jobs (although the blue-collar jobs were a minority class). There were 37 different job descriptions like Data Science, HR, Arts, Advocates etc each having their distribution. Manual filtering of some of these resumes were required as several had bad entries or were duplicates.

For the testing dataset we scraped resumes from 8 major cities– Atlanta, Austin, Boston, Chicago, New York, Raleigh, Seattle and San Fransisco. There were 1530 resumes in total.

For the data prepossessing stage, we removed-

1. URLS (eg-http://tsrtechnologyservices.com)
2. Hashtag mentions (#1Noob)
3. Extra Whitespaces
4. CCs/RTs
5. Mentions (@StrategyGuru)

Furthermore, lemmatizing and stemming was done to reduce the dimension furthers.Last for vectorising the document we used 2 different mechanisms- CountVectorizer and TFIDF. Uni-Gram and Bi-Gram words were created with a minimum document frequency of 2.The maximum number of features were fixed at 5000.

The following models with their respective accuracies were used with a 70/30 train validation split-

|  |  |  |
| --- | --- | --- |
| **Models** | **TfIdfVectorizer** | **Count Vectorizer** |
| **Logistic Regression** | 78.56% | 81.58% |
| **Naïve Bayes** | 58.12% | 73.09% |
| **Decision Tree** | 78.12% | 74.82% |
| **Random Forest** | 82.58% | 80.28% |
| **Xgboost** | 86.47% | 86.18% |
| **CAT Boost** | 88.48% | 87.48% |
| **Light GBM** | 88.48% | 86.04% |
| **ANN** | 77.84% | 76.83% |

CAT Boost gave the best results and although there were no significant differences between the above 2 scores the TFIDF models were marginally beating Count vectorizer when it came to ensemble tree-based models.