

# DocReach – Predicting Doctor Specialty using NLP

Sushil Sharma, MS, MBA | 925-968-8185 | sushilsharma71@gmail.com

## Summary: A social media marketing firm wants to target doctors/physicians based on their practice area. For example, marketing team running a campaign to target Cardiologists for a heart related news feed.

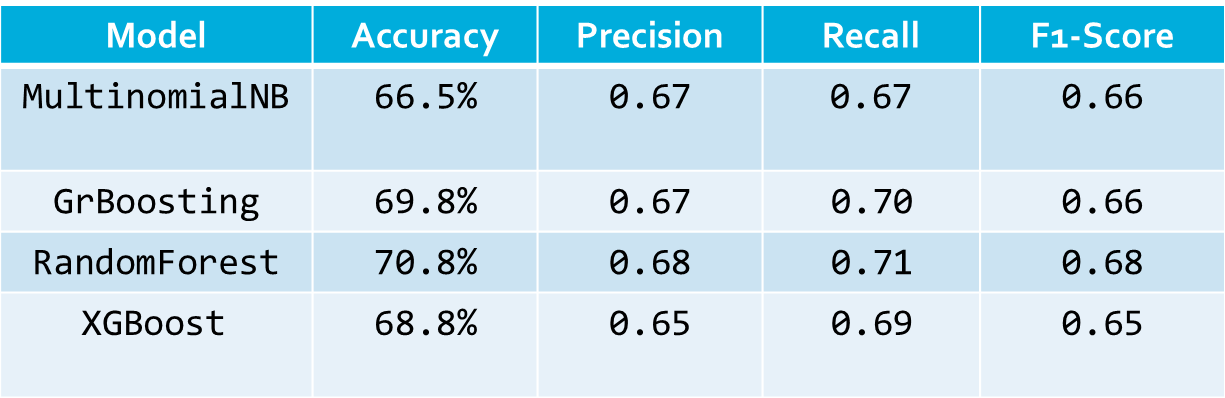
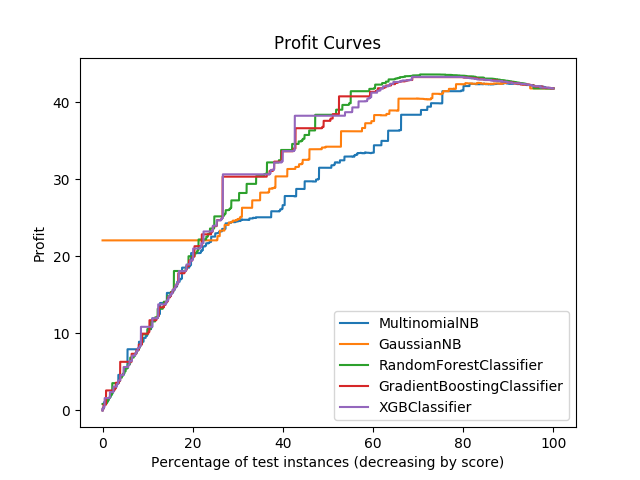
## Scope: Scope is to predict physicians under medical specialties based upon procedures performed over the past year

1. Develop a binary classifier to predict Cardiologists
2. Develop a multiclass classifier to predict top 5 Specialties

## Methodology/Process:

## Used python, pandas, and Jupyter Notebook to load and clean the data. Further, I applied Natural language processing using sklearn to convert text data into numeric features using bag of words and count vectorization. I applied multiple algorithms to build and select a classification model viz. Multinomial NaïveBayes, RandomForest, GradientBoosting, and XGBoost.

## Results:

* Received better results using TFIDF over Count-Vector
* Binary Classifier: used a cost-benefit matrix to calculate profits for each classifier for different thresholds. Random Forest performed best with a profit of $43.6, F1-Score:0.81, and accuracy: 80.5%
* Multiclass Classifier: Random Forest performed best with a F1-score of .71, and accuracy of 70.8%

## Next Steps:

* Explore additional features to improve performance
* Unsupervised and Deep learning

## Tech used:

Anaconda /Python, Pandas, Numpy, Scikit-learn, Jupyter Notebook, Spyder, AWS EC2, Ubuntu 16.04, Git, ssh, tmux

## Project links:

<https://github.com/krishnatray/galvanize-dsi-capstone>