Tanya
Seegmiller

CAPSTONE OPTION 1

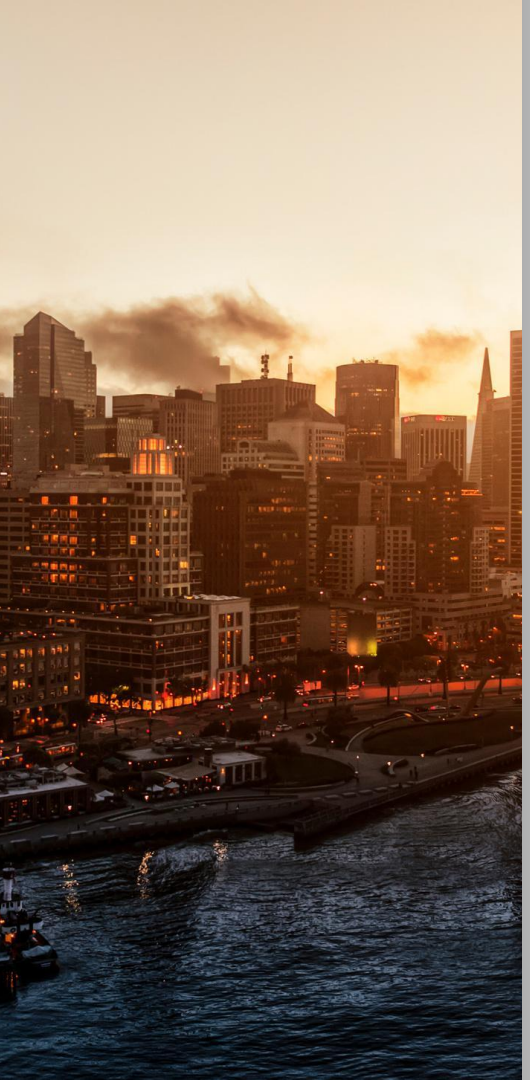
Healthcare Employee Attrition

PROBLEM STATEMENT

RN turnover costs a hospital $56,277 per nurse. The avg. annual cost of RN turnover is $4.82m per hospital. For every 20 travel RNs eliminated, the avg hospital can save $2 million.

MedicalMatch is a medtech startup committed to revolutionizing healthcare staffing. As a soon to be data scientist on their product solutions team, I am tasked with analyzing healthcare employee attrition data to quantify its financial impact on hospitals and clinics. This analysis will enable the marketing team to craft targeted campaigns promoting MedicalMatch's innovative staffing solutions.

Audience: Ali Phillips (Founder), Astrid Vreugdenhil (Data Scientist), Josh Wymer (Chief Health Information & Data Strategy Officer), Bret Lucarelli (CTO)



GOALS:

- Identify factors contributing to employee attrition.
- Quantify the financial impact of employee attrition on hospitals and clinics
- Develop predictive models for attrition prediction.

SUCCESS METRICS:

- Model will be scored on accuracy, recall, precision, and f1. The final model will perform as well as possible across all 4 metrics will as little overfitting as possible.

DATA SOURCE:

- Excel BI Analytics - IBM HR Analytics for Attrition (synthetic data)

DATA INFO:

- 50,000 rows, 35 columns
- 25 int64, 10 object

Capstone Option 2:

Disease Detection Model: Multi Classification

Tanya Seegmiller

Problem Statement

According to researchers at Johns Hopkins “each year an estimated 371,000 deaths occur from misdiagnoses, in addition to 424,000 permanent disabilities. Diagnostic errors are, by a wide margin, the most under resourced public health crisis we face.”

Doctors for Better Diagnoses, a non-profit volunteer organization looking to improve patient outcomes has tasked me with building a model that will assess patient symptoms and make accurate prognosis predictions.

Potential Audience: Everyone

Accurate infection and disease detection affects all of us.

Goals:

1) Develop and train a classification model that will accurately predict the diagnoses in my labeled data.
2) Create a model that will help reduce the number of misdiagnosed infections/diseases.

Success Metrics:

I will be scoring the model on accuracy, recall, precision and f1 scores. The best model will perform as well as possible across all 4 metrics with as little overfitting as possible.

Data Source and Details:

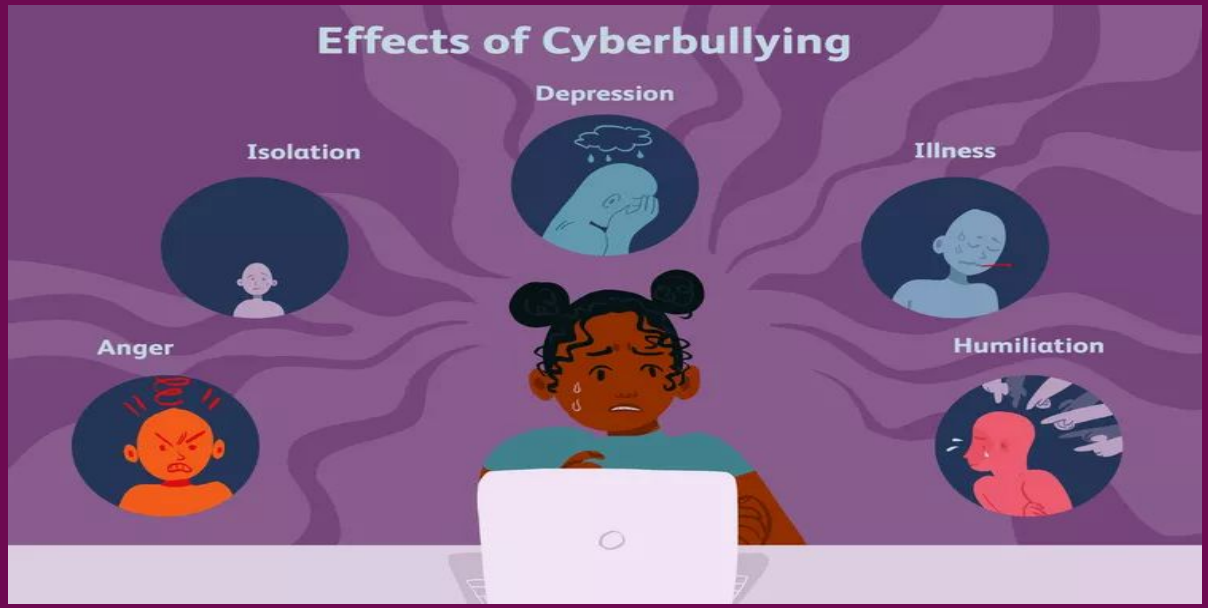
Source: Kaggle

Details:

- **2 files:**
  - Train.csv 4920 Rows, 133 columns (132 features/symptoms, 1 target/prognosis)
  - Test.csv 42 Rows, 133 columns (132 features/symptoms, 1 target/prognosis)
- The symptoms are mapped to 42 diseases

Capstone Option 3:

Cyberbullying Detection Model: NLP Multi Classification



Tanya Seegmiller

Problem Statement:

"36.5% of middle and high school students have felt cyberbullied and 87% have observed it. It's an ever increasing problem as social media usage increases across all age groups."

The Global Cyber Safety Alliance is an advocacy group dedicated to improving cyber safety and cyberbullying through data driven initiatives and collaboration across communities and countries. I have been hired to create a NLP model that will detect bullying and label the class.

Potential Audience

Non Profit Groups, Law Enforcement, Schools, Mental Health Professionals, etc.

Cyberbullying NLP Classification

Goals:

1) Develop and train a NLP classification model that will accurately identify cyberbullying tweets.
2) Bring awareness to the prevalence and severity of cyberbullying and its adverse effects.

Success Metrics:

Model will be scored on accuracy, recall, precision and f1. The best model will perform as well as possible across all 4 metrics with as little overfitting as possible.

Data Source:

Kaggle

Data Details:

47,000 tweets labelled according to the class of bullying:

- **Age: 8,000**
- **Ethnicity: 8,000**
- **Gender: 8,000**
- **Religion: 8,000**
- **Other: 8,000**
- **Not cyberbullying: 8,000**

Trigger Warning:

The tweets in this dataset either describe a bullying event or are the bullying event themselves.

Citation

J. Wang, K. Fu, C.T. Lu, “SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection,” Proceedings of the 2020 IEEE International Conference on Big Data (IEEE BigData 2020), December 10-13, 2020.