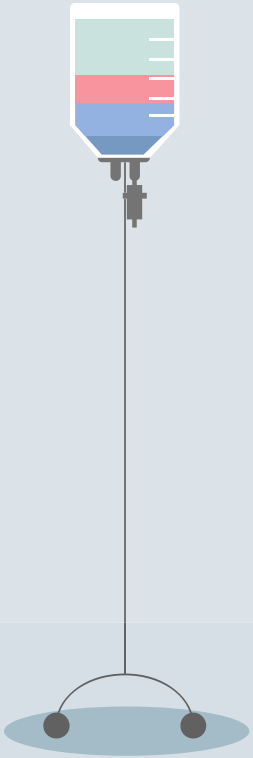
Inpatient Length of Stay

By: Saif Rahman

AGENDA



01
INDUSTRY INTRODUCTION

02
DATASET

03
DATA PRE-PROCESSING

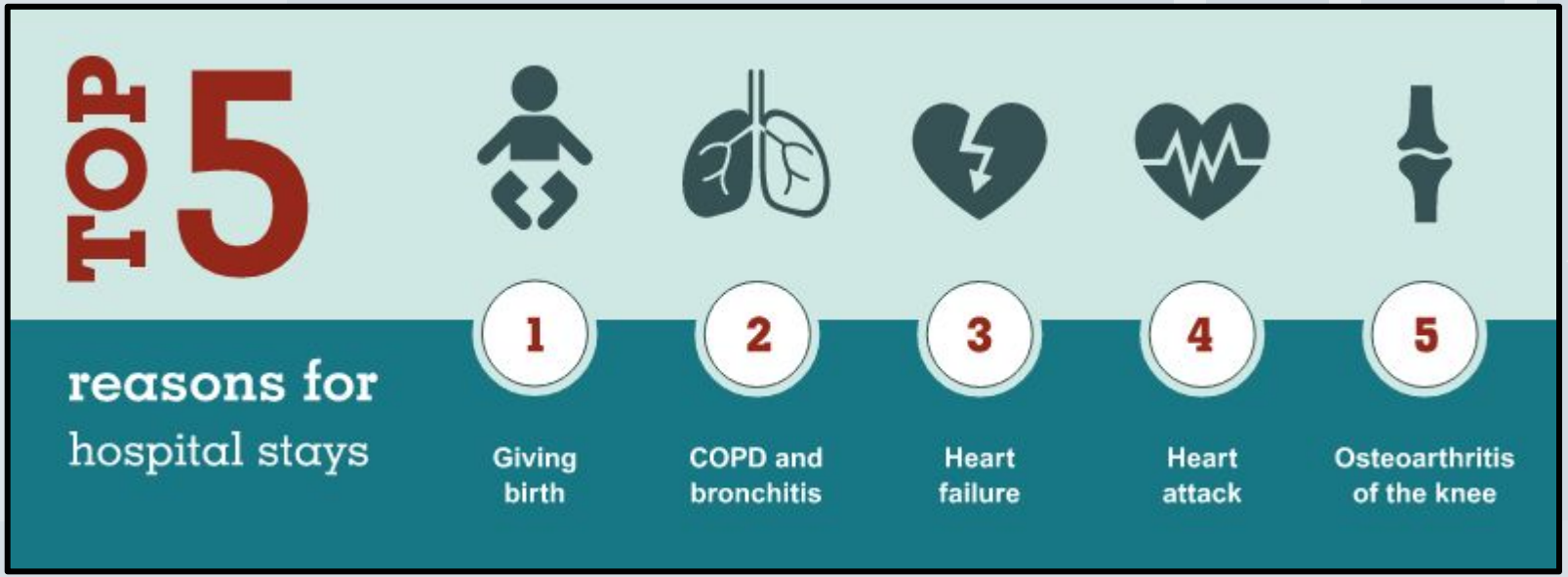
04
MODELLING

05
NEXT STEPS

INDUSTRY: HEALTHCARE

- In 2018, the US had over 36.3 million hospital admissions
- Large amounts of patient information
- Strong need for data scientists to improve patient care quality

INDUSTRY INTRODUCTION



DATASET

New York Hospital Dataset

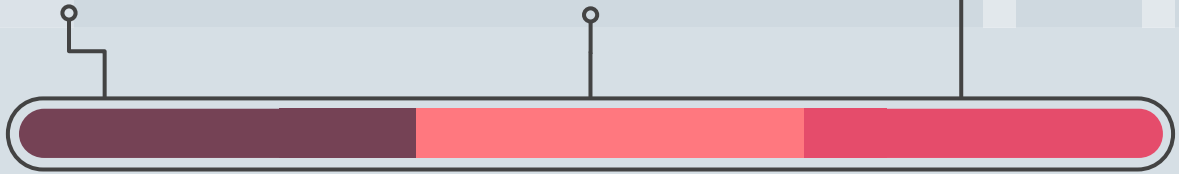
- 2,346,931 records
- 34 features
- 22 categorical columns

AGE: 25 - 35
GENDER: Female
RISK OF MORTALITY: Minor
LENGTH OF STAY: 4 Days

Disease and Disorders of the Respiratory System

Electrocardiogram

Short-term Hospital



APR MDC Description

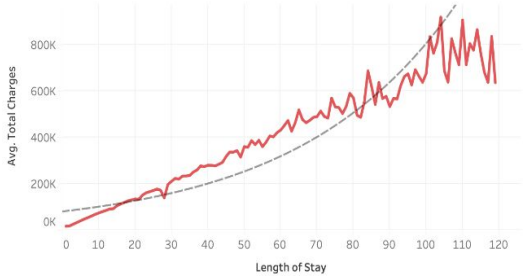
CSS Proc Description

Patient Disposition

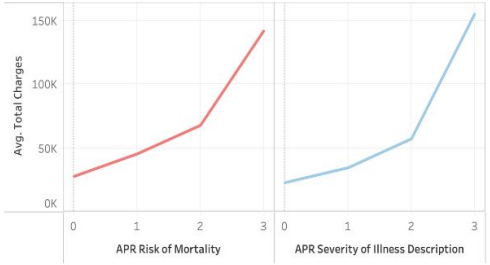
DATASET: EXPLORATION

Hospital Data Exploration

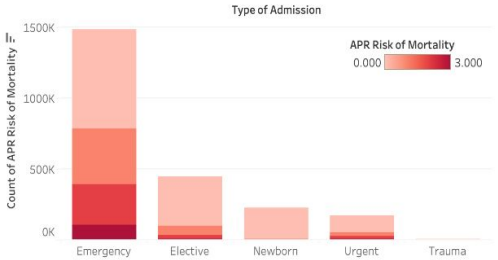
Average Charge vs LOS



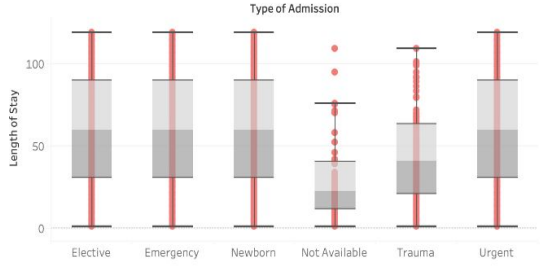
Average Charges vs APRs



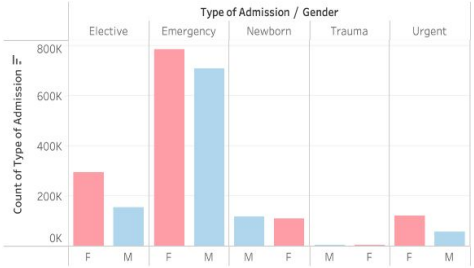
APR Counts of Admissions



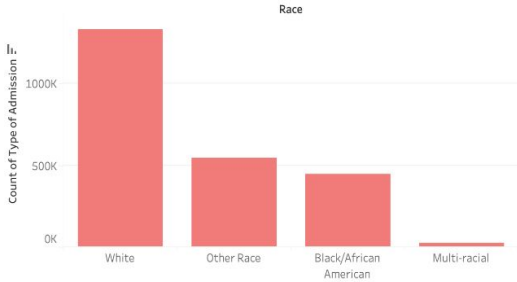
Admission Boxplots



Admissions Counts by Gender



Admission Counts by Race



Click here for Tableau Link

DATA PRE-PROCESSING

FILTERED

```
data1['Patient Disposition'].value_counts()
```

Home or Self Care	1572079
Home w/ Home Health Services	304373
Skilled Nursing Home	224088
Expired	51020
Left Against Medical Advice	47065
Inpatient Rehabilitation Facility	44544
Short-term Hospital	40553
Hospice - Medical Facility	12666
Psychiatric Hospital or Unit of Hosp	12096
Hospice - Home	10742
Another Type Not Listed	8351
Facility w/ Custodial/Supportive Care	6680
Court/Law Enforcement	3887
Medicare Cert Long Term Care Hospital	3445
Cancer Center or Children's Hospital	2906
Hosp Basd Medicare Approved Swing Bed	1550
Federal Health Care Facility	621
Critical Access Hospital	153
Medicaid Cert Nursing Facility	112

Name: Patient Disposition, dtype: int64

=

Home	1857962
Other	331028
Facility	63906
Hospital	58219

DATA PRE-PROCESSING

ENCODED

```
data1['Hospital County'].unique()
```

```
array(['Allegany', nan, 'Cattaraugus', 'Chautauqua', 'Erie', 'Niagara',
       'Genesee', 'Chemung', 'Orleans', 'Wyoming', 'Monroe', 'Albany',
       'Livingston', 'Ontario', 'Wayne', 'Schuyler', 'Steuben', 'Cayuga',
       'Yates', 'Oswego', 'Broome', 'Cortland', 'Jefferson', 'Oneida',
       'Chenango', 'Herkimer', 'Onondaga', 'Madison', 'Lewis',
       'St Lawrence', 'Tompkins', 'Columbia', 'Fulton', 'Montgomery',
       'Otsego', 'Rensselaer', 'Saratoga', 'Clinton', 'Schenectady',
       'Delaware', 'Franklin', 'Essex', 'Warren', 'Dutchess', 'Orange',
       'Schoharie', 'Putnam', 'Rockland', 'Sullivan', 'Ulster',
       'Westchester', 'Bronx', 'Kings', 'Manhattan', 'Queens', 'Richmond',
       'Suffolk', 'Nassau'], dtype=object)
```

=

```
array([0.0008961 , 0.0025373 , 0.00446581, 0.05177328, 0.00731552,
       0.0017948 , 0.00591922, 0.00052096, 0.00094803, 0.0450843 ,
       0.02810159, 0.00097053, 0.00507764, 0.00239927, 0.00025529,
       0.00268052, 0.00228937, 0.00030678, 0.00199514, 0.01222484,
       0.00165202, 0.00463283, 0.01277306, 0.00069447, 0.00031111,
       0.03354528, 0.0019973 , 0.00060447, 0.00474663, 0.00307903,
       0.0024737 , 0.00124442, 0.00293235, 0.0052122 , 0.00511225,
       0.00393663, 0.00421615, 0.00953133, 0.00027649, 0.00195274,
       0.00020034, 0.00603908, 0.01394651, 0.01681872, 0.00019947,
       0.00289211, 0.01362286, 0.00183634, 0.00481586, 0.05064222,
       0.07979482, 0.10617775, 0.1704225 , 0.08432943, 0.02433241,
       0.06781575, 0.07763309])
```

DATA PRE-PROCESSING

ENGINEERED

```
data['num_typologies'] = \
data[
    ['Payment Typology 1',
     'Payment Typology 2',\
     'Payment Typology 3']]\
.notnull().sum(axis=1)
```

=

```
data['num_typologies'].value_counts()

2    918736
1    725011
3    667368
Name: num_typologies, dtype: int64
```

DATA PREPROCESSING

FEATURE SELECTION

```
drop_features = ['Payment Typology 1','Payment Typology 2','Payment Typology 3','Total Charges','Length of Stay']
```

```
X = pd.get_dummies(data.drop(columns=drop_features),drop_first=True)
X = X.rename(columns=col_renamer(X.columns)) # apply column renamer
y = data['Length of Stay']
```

```
X.shape
```

```
(2311115, 59)
```

MODELLING: METRICS

$$R^2 = 1 - \frac{SSE}{SST}$$

A statistical measure of how close the data are to the fitted regression line, 'goodness of fit`. Representing a value from 0 to 1.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Indicates the absolute fit of the model to the data; how close the observed data points are to the model's predicted values.

MODELLING: MACHINE LEARNING

Linear Regression

R2 Score: 0.231
RMSE: 6.962

Ridge Regression

R2 Score: 0.231
RMSE: 6.962

Lasso Regression

R2 Score: 0.23
RMSE: 6.965

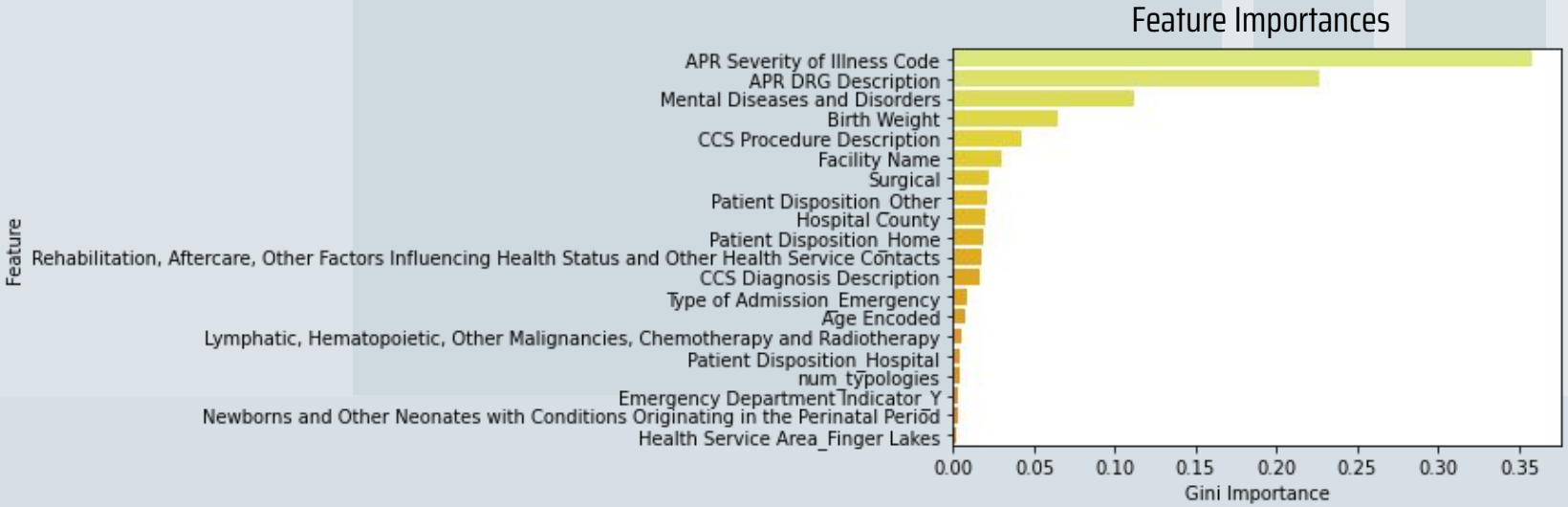
Decision Tree Regression

R2 Score: 0.384
RMSE: 6.231

R^2 of **0.384** indicates that 38.4% of the variability in the outcome data cannot be explained by the model

MODELLING: Decision Tree

- Pipeline implementation using StandardScaler()
- Hypertuned using grid search → Optimal tree with max_depth of 10



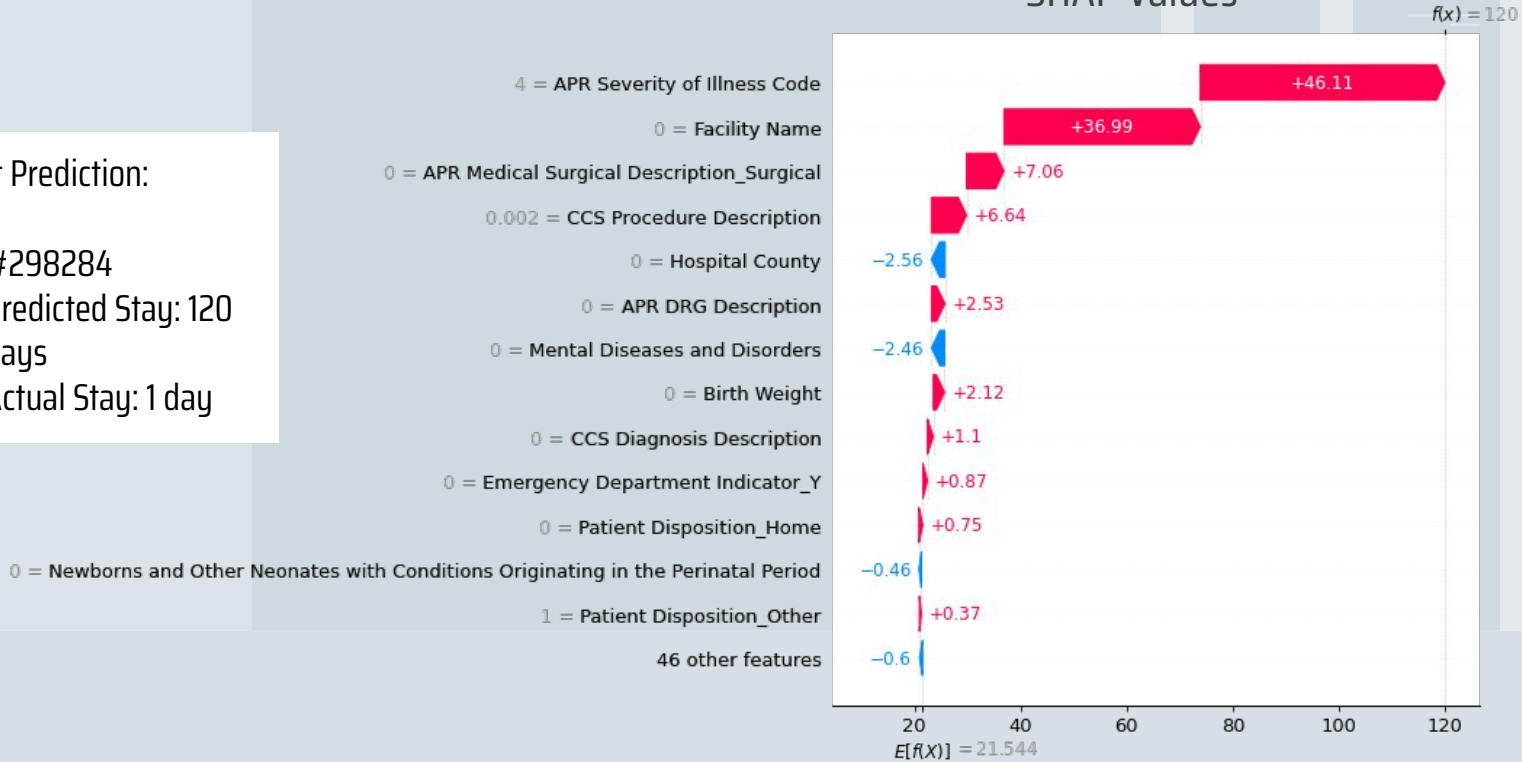
MODELLING: DECISION TREE

Incorrect Prediction:

Patient #298284

- Predicted Stay: 120 days
- Actual Stay: 1 day

SHAP Values



MODELLING: APPLICATION

- Predicting length of stay for inpatient organization and scheduling
- Determining amount of resources that must be allocated for patients
- Key metric for predicting total charges
- Additional information for patients to evaluate their situation

Next Steps

- Additional patient information
- Better feature engineering
- Dealing with categorical columns
- Implementing a neural network

THANKS

Does anyone have any questions?

Github Link: