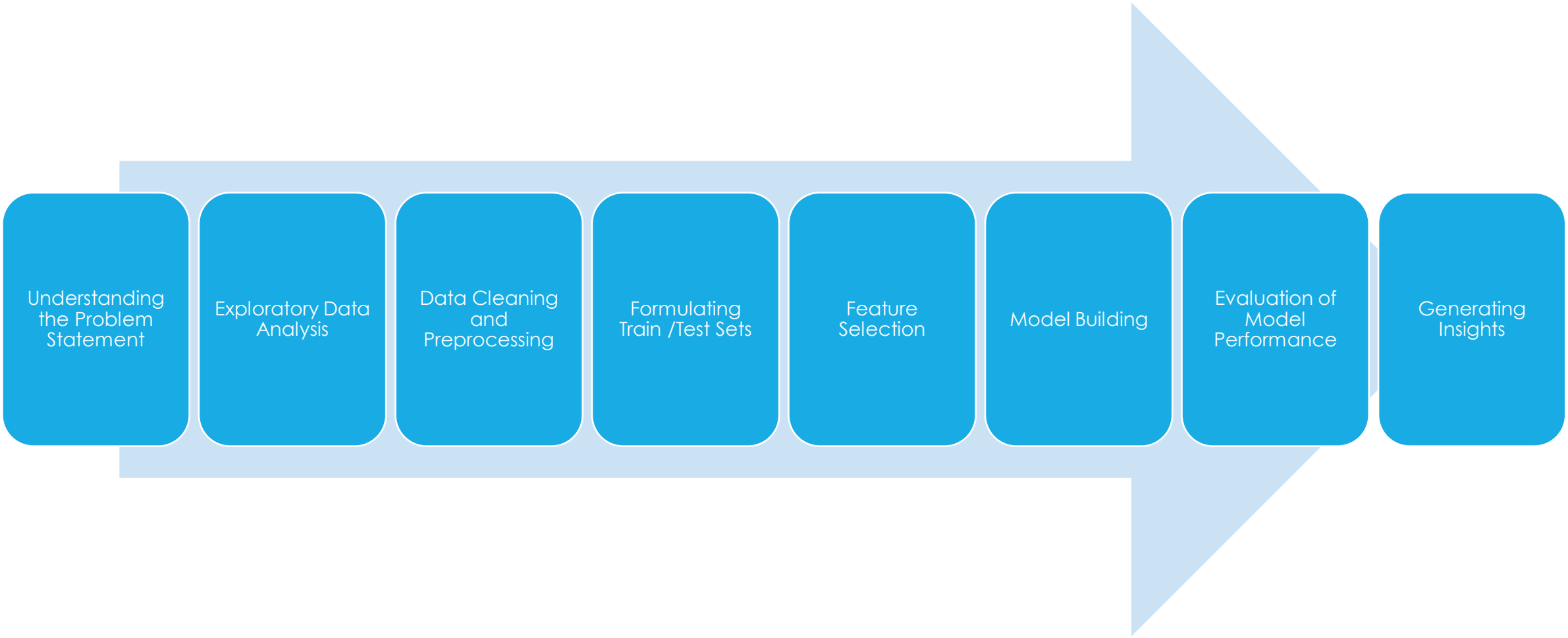
Predicting Patient Length Of Stay

Project part of Data Science Hackathon 2018 at Saint Peter's University conducted by University Of South Florida, Department Of Surgery

Pranay Katta

Process Flow



Problem Statement

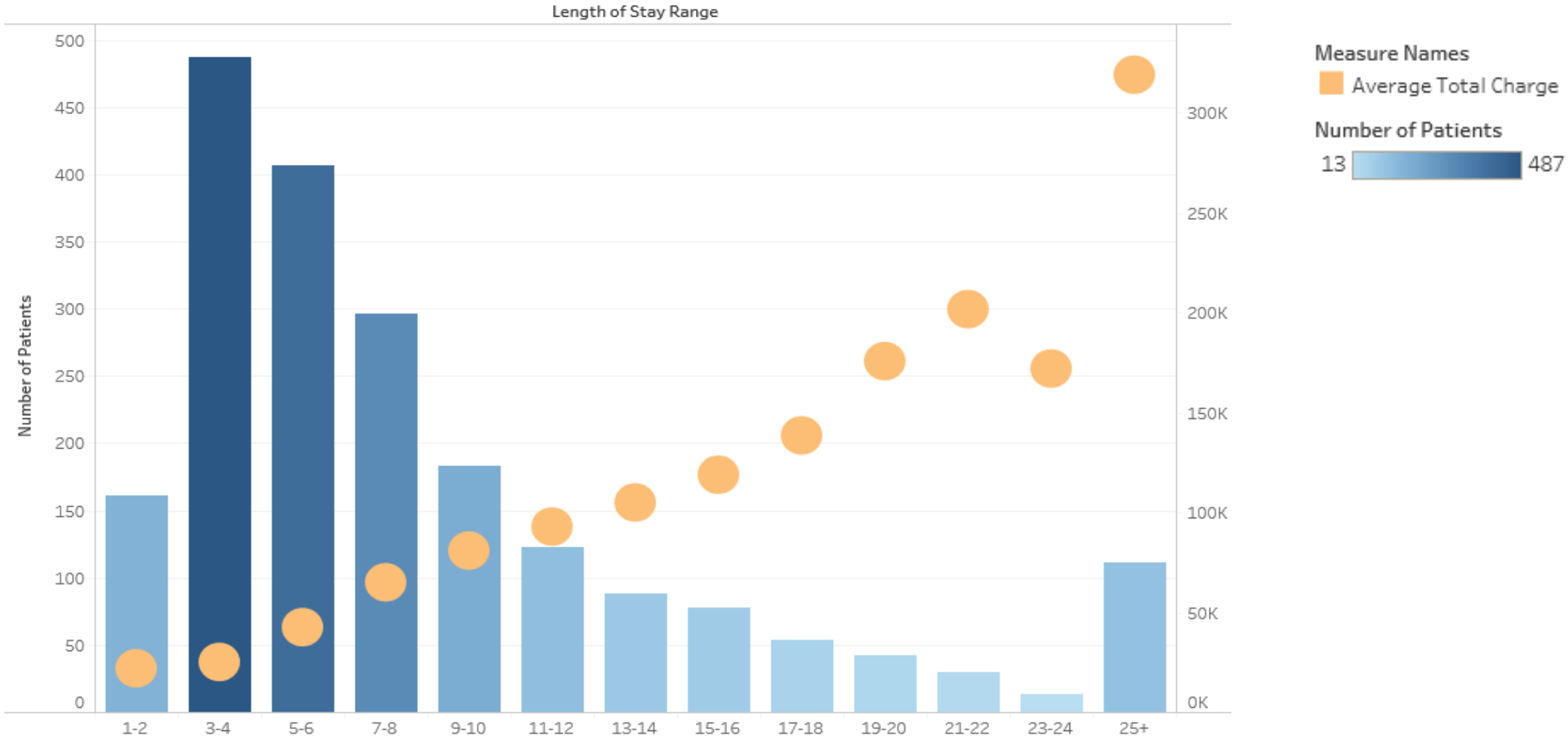
- Predicting Ideal Length Of Stay of Liver Cohorts who got readmitted within first 30 days of their index admission

  - Reduces costs to patients, hospitals and insurance companies
  - Increase quality of treatment to other patients

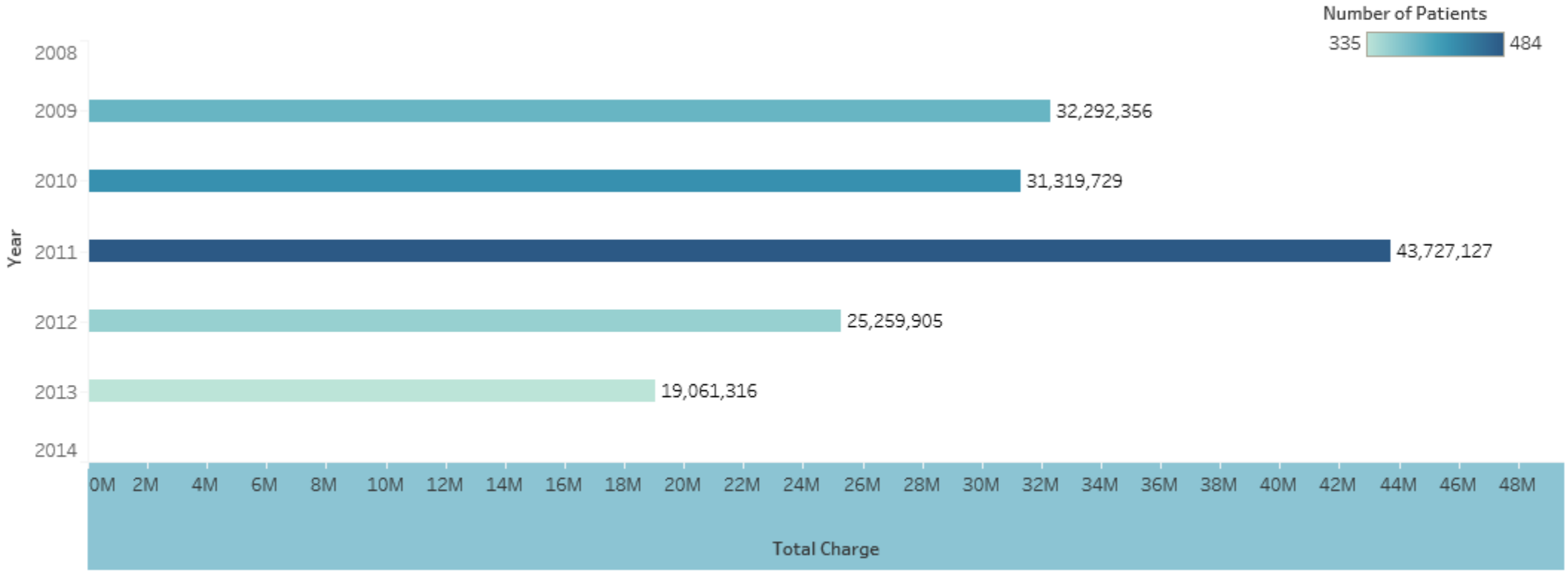
Dataset

- (HCUP) Health Care Cost and Utilization Agreement dataset 2009 - 2013
- 27,607 Observations
- 114 Independent Variables
  - 16 Continuous Variables
  - 98 Categorical Variables
- Dependent Variable – (LOS) Length Of Stay

Exploratory Data Analysis



Exploratory Data Analysis



Understanding the data

GOOD VISITS

LOS	READMIT	VisitLink
3	0	1671
6	0	1671
2	0	1671
2	0	1671
7	0	2577
7	0	699
4	0	699
4	0	191943
6	0	191943

READMISSION VISITS

INDEX

LOS	READMIT	VisitLink
2	0	753922
8	1	753922
5	1	753922
5	1	753922
6	0	191943
1	1	191943
18	1	191943
4	0	191943
6	0	191943

BAD VISITS

LOS	READMIT	VisitLink
4	1	4688420
1	0	4688420
2	0	4688420
14	1	12707
6	1	16131

Data cleaning and pre-processing

- Check Point 0

```
In [5]: print(df.isnull().sum())
ATYPE          0
AWEEKEND       0
BDTOT          0
CHC            0
CM_AIDS     1773
CM_ALCOHOL  1403
CM_ANEMDEF     0
CM_ARTH        0
CM_BLDLOSS     0
CM_CHF         0
CM_CHRNLUNG    0
CM_COAG        0
CM_DEPRESS  1403
CM_DM          0
CM_DMCX        0
CM_DRUG     1403
CM_HTN_C       0
CM_HYPOTHY     0
CM_LIVER    1403
CM_LYMPH       0
```

	ADC	ADMTOT	AGE	BDTOT	FTMDTF	FTRES	FTRNTF	NCHRONIC	NDX	NPR	SUF
count	31283.000000	31283.000000	31283.000000	31283.000000	31283.000000	31283.000000	31283.000000	31283.000000	31283.000000	31283.000000	31283.0
mean	504.913371	32084.854426	58.755171	645.792827	174.708979	318.544577	1307.326695	4.402647	9.476009	3.386152	10482.9
std	413.937023	25137.223522	14.740317	508.495655	343.847205	407.436866	1117.885653	2.681588	5.252218	3.077413	8057.3
min	3.000000	192.000000	18.000000	16.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.0
25%	224.000000	16327.000000	50.000000	315.000000	2.000000	4.000000	487.000000	2.000000	6.000000	1.000000	4329.0
50%	386.000000	24515.000000	60.000000	470.000000	23.000000	139.000000	994.000000	4.000000	9.000000	3.000000	8303.0
75%	698.000000	39957.000000	69.000000	806.000000	148.000000	476.000000	1847.000000	6.000000	13.000000	5.000000	13513.0
max	1952.000000	130100.000000	112.000000	2338.000000	1907.000000	1688.000000	5284.000000	18.000000	45.000000	31.000000	49829.0

1. Exploring Five-number summary of variables.
2. Imputing Missing Values:
   A) For Comorbidities Measure: Default to 0
   B) For Multivalued Categorical: Default to ‘Others’
   C) For Binary Categories: Random(0,1)
   D) For Continuous: Mean
   E) Remove if an entire feature has no contribution to the data
3. Total Selected Features after pre-processing – 38
4. Creating dummy variables for non binary

Approach

Check Point 1

- Building a model using patients data who did not revisit hospital within 30 day window and using the parameters to estimate ideal length of stay for the patients who visited the hospital within 30 day window

  OR

- Building a model with both types of patient visits and determining ideal length of stay for the index visits

Formulating Train / Test Sets

Check Point 2

- Assuming each visit is independent of the other
- Dividing the data into train, validation and test sets in ratio (7:1:2)

  - # of Unique Patient revisits within 30 days window – 3137
  - # of Unique Patient revisits after 30 days window – 13,934
  - Test set contains data of all the index visits of patients who are readmitted within 30 days of index visit and we are interested in estimating ideal LOS for those patients.

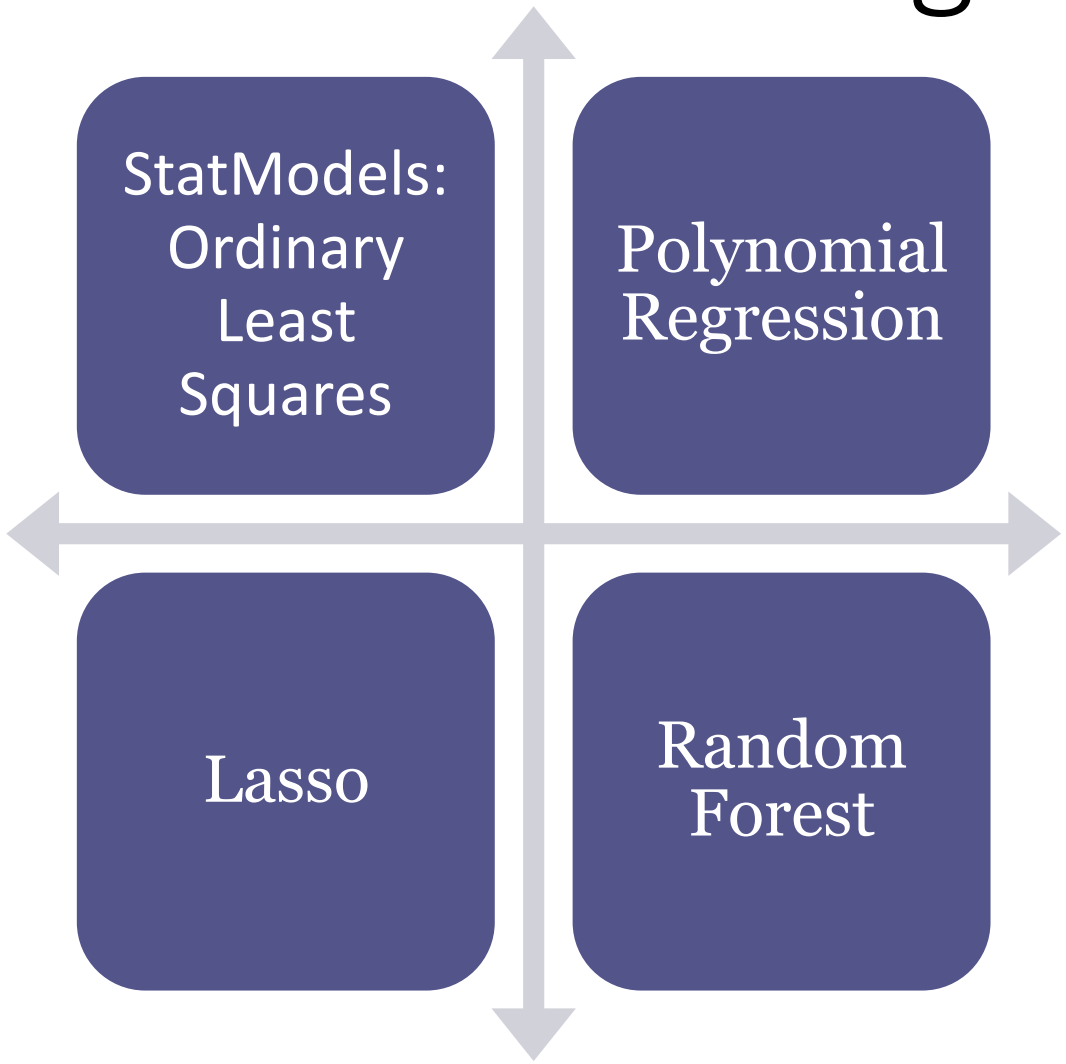
Feature Selection

Check Point 3

- Important Features : **Post Operative Problems, Comorbidity Measures, Disposition of Patient at Discharge (Others)** and **Age**
- High correlations are found between ADC (Average daily Census), ADMTOT (Total Facility Admissions), BDTOT (Total Beds Staffed)

Data modeling

- Check Point 4



Evaluating Model Performance

Ordinary Least Squares

	Fold1	Fold2	Fold3	Fold4	Fold5	Average
Adjusted R Sq.	0.698918	0.700705	0.653858	0.632295	0.665496	**0.670254**
RMSE	4.45	4.46	4.165	4.028	4.239	**4.2684**

Testing Data

RMSE	4.3864
MSE	19.2405
Abs Mean	2.34035

Lasso

	Fold1	Fold2	Fold3	Fold4	Fold5	Average
Adjusted R Sq.	0.807	0.808	0.808	0.815	0.809	**0.8094**
RMSE	4.27	4.006	4.43	4.17	4.23	**4.2212**

Testing Data

RMSE	4.19923
MSE	17.63353
Abs Mean	2.225504

Random Forest

RMSE	5.809707
MSE	33.75269
Abs Mean	3.432953

Adj R sq Train	0.4886
Adj R sq Test	0.5518

Polynomial Regression (degree = 2)

RMSE	4.68625
MSE	21.96094
Abs Mean	2.83315

Adj R sq Train	0.622
Adj R sq Test	0.591

Calculation Of Cost Saved

VISITLINK	ACTUAL LOS	IDEAL LOS	DIFFERENCE	TOTAL CHARGES	TOTAL CHARGES/DAY	TOTAL_CHGS READMISSION	INCREMENTAL TOTAL CHARGES	POTENTIAL SAVINGS
1858	2	5.27640628	3.276406288	11021.15	5510.575195		18054.88281	-10889.78
1858	1		3.276406288	7165.1		7165.1		
2707	3	7.5359044C	4.535904408	14125.76	4708.586426		21357.69727	-10830.25
2707	5		4.535904408	10527.45		10527.45		
5598	16	19.9975051	3.997505188	141957.12	8872.320313		35467.14453	809.625
5598	9		3.997505188	36276.77		36276.77		
72031	13	17.8760948	4.876094818	324876.67	24990.51367		121856.1172	137705.1
72031	38		4.876094818	259561.24		259561.2		
80463	10	18.4629974	8.462997437	72611.66	7261.166016		61451.23047	4674.973
80463	7		8.462997437	66126.2		66126.2		
82805	6	11.251358C	5.251358032	73824.4	12304.06641		64613.05859	-32345.71
82805	2		5.251358032	17281.35		32267.35		
82805	4		5.251358032	14986				

Findings

- **Average increase in LOS for index VisitLinks = 2.95 ~ 3 days**
- **Average Cost Saved per VisitLink = $56,523.94**

  - By increasing average LOS by 2.95 or 3 days in first index visit $56,523.94 costs will be saved on an average per VisitLink.
  - By staying 3 more days on the index visit, 3 extra days or 6 days in total of stay during readmission can be eliminated on an average

Other Approaches

- Building Model using only the ideal data and estimating the LOS for non-ideal data
- Logistic Regression – Assuming each visit is independent
- Markov Models – Assuming each visit is not independent


Q
A