



Forecasting Patient Enrolment for Clinical Trials

5th Team Project Sprint Review
June 19th, 2020



01

Feature Engineering

Raw, newly generated and additionally crawled features

02

Preprocessing

Different preprocessing techniques necessary for the model

03

Analysis

Interesting findings on our clinical trials

04

Model

Using regression or transforming it into a classification problem



Feature Engineering

- Raw Features
- Additionally crawled features
- Newly Generated Features
 - Simple Ones
 - Complex Ones

Raw Features

Organisational Information

- OrgFullName, OrgClass, StudyType
- LeadSponsorName, LeadSponsorClass, CollaboratorName, CollaboratorClass

Disease related features

- Condition, ConditionBrowseLeafName and ConditionBrowseLeafRelevance
- ConditionAncestorId/ConditionAncestorTerm
- ConditionMeshId/ConditionMeshTerm

Study Design related features

- EligibilityCriteria, HealthyVolunteers, Gender, StdAge
- DesignAllocation, DesignInterventionModel, DesignPrimaryPurpose
- InterventionName
- IsFDARegulatedDrug
- Enrollmentcount

Study Location

- LocationFacility, LocationCity, LocationCountry

Other Features

- ArmGroupLabel
- Keyword

Additionally Crawled Features

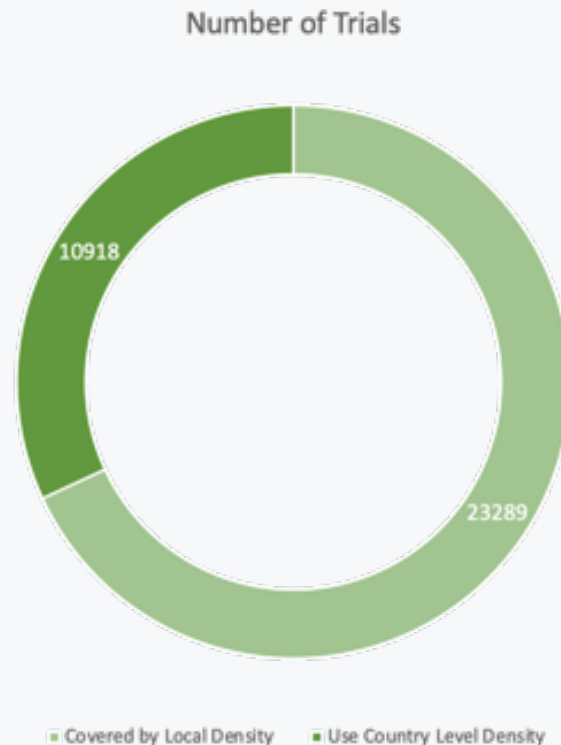


OECD Population Density Data



Local Density

- Match region with LocationState and LocationCity
- Not all trials can be covered
- ~two third of the trials
- Use the country level density for those



New (Simple) Generated Features

- Number of diseases
- Number of ConditionAncestor
- Number of Collaborators
- Length of String in EligibilityCriteria
- Number of ArmGroupLabels (of parallel groups)
- Number of LocationFacilities
- Number of Cities
- Number of Countries
- Start Year

New Generated Features

- ConditionBrowseLeafName combined with ConditionBrowseLeafRelevance
 - Number of Patients per Location Facility
 - Number of Patients per Country
 - Number of Patients per Country according to size of country
-
- Number of Patients
 - Number of Patients/(Sites * Months)
 - Number of Patients * Sites

New Generated Features - Country and Hospital Data

- Avg of population of countries involved
 - Avg of lifeExpectancy of countries involved
 - Avg of GDP of countries involved
 - Avg of density of countries involved
 - Avg of migrantsNet of countries involved
 - Avg of sizeInKm2 of countries involved
 - Avg of urbanPopulation of countries involved
-
- Avg of hospitalBeds of hospitals involved

New Generated Features - Treatment duration

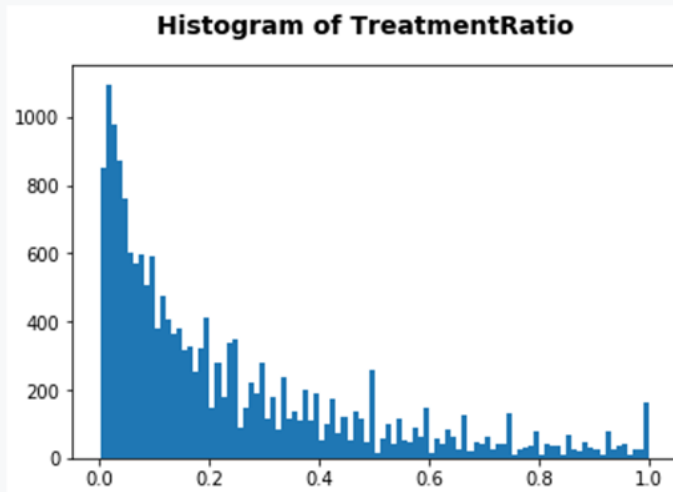
- Two features can be used to estimate the treatment duration
 - OutcomeMeasureTimeFrame
 - PrimaryOutcomeTimeFrame
- Problems:
 - 60% Missing in OutcomeMeasureTimeFrame, 10% in PrimaryOutcomeTimeFrame
 - Input = Free Text without structure
 - Some rows mention a “baseline” duration without specification of the baseline
 - Terms like “weekly”, “2-4 years”, “2 months and then biweekly until ...” hard to detect and interpret automatically

New Generated Features - Treatment duration

- Further analysis:
 - Create Feature “TreatmentRatio” to get an estimate of the average time the treatment of a patient takes from the overall enrollment duration
 - Some detected values are larger than the total enrollment duration

Mean Ratio with outliers: 0.273

Mean Ratio without outliers: 0.232



New Generated Features - Treatment duration

- Approach:
 - applied some natural language processing techniques to preprocess the text
 - i.e. split string into word tokens, to lower case, create numbers from string representation
 - look for patterns in data like “week”, “day”, “month”, “year”
 - based on those words find numeric value before that term
 - if necessary, modify values
 - for ranges (i.e. 2-4) take upper bound
 - for decimals (i.e. 3.4) take rounded values
 - multiply or divide values to get number of months (i.e. year → multiply by 12)
 - create new feature “TreatmentDuration” for rows where a duration could be detected

Preprocessing

- Preprocessing Techniques
 - One Hot Encoding
 - Ordinal Encoding
 - Normalization
- Outlier Detection



Encoder

- StudyType
- LeadSponsorClass
- HealthyVolunteers
- IsFDARegulatedDrug

	NCTId	StudyType	LeadSponsorClass	EnrollmentCount	HealthyVolunteers	IsFDARegulatedDrug	EnrollmentDuration	S
0	NCT00000143	0.0	5.0	61	1.0	1.0	37	
2	NCT00000177	0.0	4.0	120	1.0	1.0	39	
3	NCT00000200	0.0	4.0	19	1.0	1.0	12	
6	NCT00000261	0.0	5.0	14	1.0	1.0	11	

One Hot Encoding

- CollaboratorName
- CollaboratorClass
- Gender
- StdAge
- OrgClass

Open for discussion:

- Condition/ConditionMeshTerm (or better: ConditionAncestorTerm)
- Country
- Keywords (Attribute 'Keyword' itself or extracted from 'BriefTitle' 'BriefSummary', 'description', 'EligibilityCriteria' or 'FlowRecruitmentDetails')

Normalization

#DiffCities	#DiffCountries	#Pts/Sites	#Pts/Country	AvgPop	AvgLifeExpec	AvgGDP	AvgDensity	AvgMigr	AvgSize	AvgUrbPop
16.0	1.0	3.210526	61.0	331002656.0	78.0	2.054434e+13	36.0	954806.0	9147420.0	83.0
1.0	1.0	168.000000	168.0	331002656.0	78.0	2.054434e+13	36.0	954806.0	9147420.0	83.0
4.0	1.0	94.750000	379.0	331002656.0	78.0	2.054434e+13	36.0	954806.0	9147420.0	83.0
6.0	1.0	72.000000	432.0	331002656.0	78.0	2.054434e+13	36.0	954806.0	9147420.0	83.0





#DiffCities	#DiffCountries	#Pts/Sites	#Pts/Country	AvgPop	AvgLifeExpec	AvgGDP	AvgDensity	AvgMigr	AvgSize	AvgUrbPop
16.0	1.0	3.210526	61.0	0.312591	0.076823	1.5009	-0.696615	1.439702	1.09517	0.423373
1.0	1.0	168.000000	168.0	0.312591	0.076823	1.5009	-0.696615	1.439702	1.09517	0.423373
4.0	1.0	94.750000	379.0	0.312591	0.076823	1.5009	-0.696615	1.439702	1.09517	0.423373
6.0	1.0	72.000000	432.0	0.312591	0.076823	1.5009	-0.696615	1.439702	1.09517	0.423373

Outlier detection

2 approaches:

- Based on Interquartile Range: outliers are all values outside $[Q1 - 1.5 * IQR; Q3 + 1.5 * IQR]$
- Based on data distribution: outliers are all values below 5th or above 95th percentiles

Also based on typical expectations

PHASE 0	PHASE 1	PHASE 2	PHASE 3	PHASE 4
Lab Studies	Safety and dosage	Efficacy and side effects	Efficacy and monitoring of adverse reactions	Safety and efficacy
				
Typical Number of Participants	20 to 100	Up to several hundred	300 to 3000	Several thousand
Typical Duration	Several months	Up to 2 years	1 to 4 years	

```

outlierSettings = {
  'Phase 1': {
    'EnrollmentCount': [9, 95.5],
    'EnrollmentDuration': [1, 84.4]
  },
  'Phase 2': {
    'EnrollmentCount': [11, 363.25],
    'EnrollmentDuration': [5, 101]
  },
  'Phase 3': {
    'EnrollmentCount': [26, 1584],
    'EnrollmentDuration': [5, 89]
  },
  'Phase 4': {
    'EnrollmentCount': [12, 956.8],
    'EnrollmentDuration': [6, 81.4]
  }
}

```

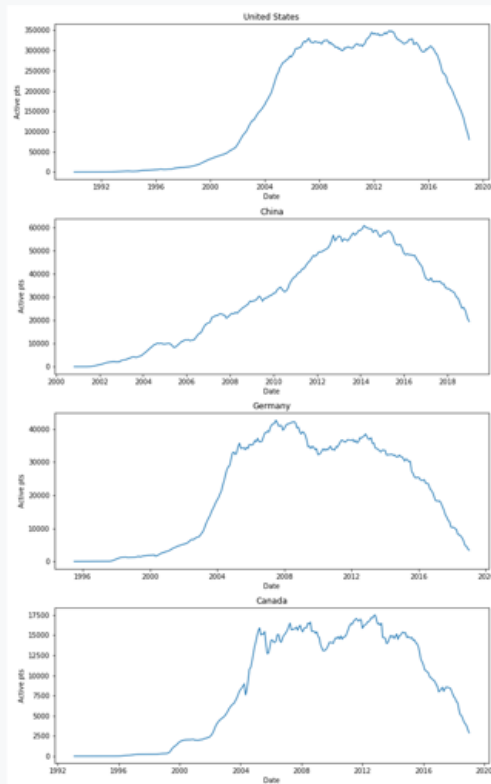
Analysis

- Plotting Enrollment Count per Country with the enrollment distributed over time

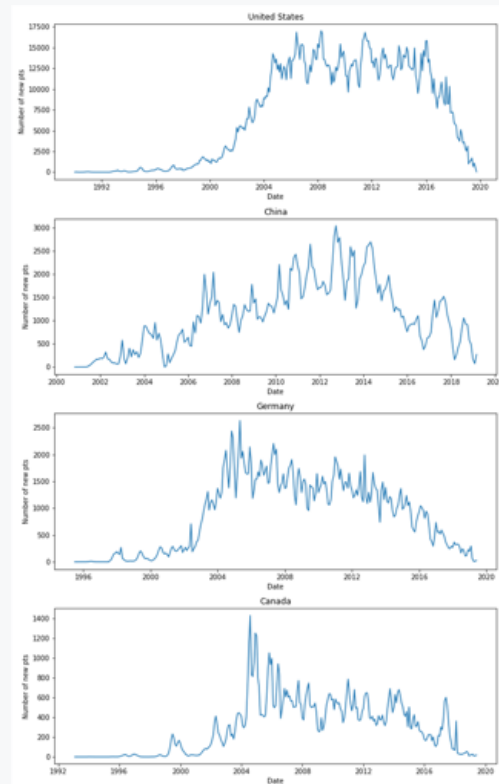
Patients per country time-series

- Analyzing on only one-phase studies: 32,999 studies
- For multiple-countries studies: distribute number of patients based on World Share factor
- New assumption: enrollment duration take about 30%* of study duration timeline
- Distributing number of patients to the first 30% of study duration and using logistic function

Active patients per country



New patients per country



Models

- Regression Models
- Regression → Classification

Features

- OrgFullName, **OrgClass**
- **Condition**, ConditionBrowseLeafName and ConditionBrowseLeafRelevance
- ConditionAncestorId/ConditionAncestorTerm
- ConditionMeshId/ConditionMeshTerm
- LeadSponsorName, **LeadSponsorClass**, CollaboratorName, **CollaboratorClass**
- **EligibilityCriteria**, **HealthyVolunteers**, **Gender**, **StdAge**
- **LocationFacility**, **LocationCity**, **LocationCountry**
- InterventionName
- **IsFDARegulatedDrug**
- **ArmGroupLabel**
- Keyword
- DesignAllocation, DesignInterventionModel, DesignPrimaryPurpose
- **Enrollmentcount**
- **StartYear**

Generated Features

- **LengthEligi**
- **#Condition**
- **#Collabs**
- **AvgPop**
- **AvgLifeExpec**
- **AvgGDP**
- **AvgDensity**
- **AvgMigr**
- **AvgSize**
- **AvgUrbPop**

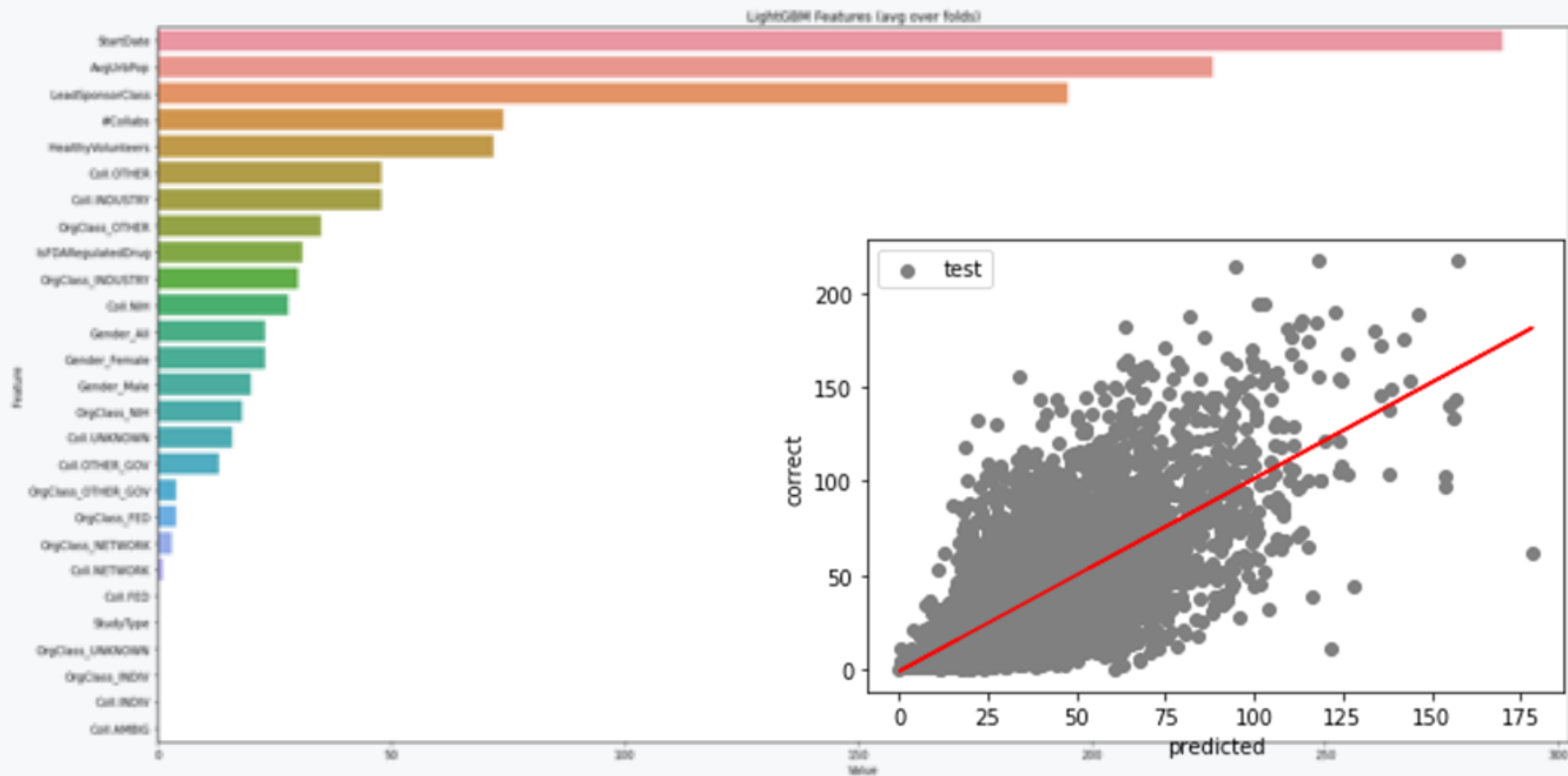
Data

- **Target:** EnrollmentDuration
- **Excluded Phases:** Early Phase 1, Phase 1
- **Number of trials:** ~ 20 000
- **Training Data:** 2/3
- **Feature Selection:** Backward elimination

Regression

Approach	R^2	MAE	MSE	RMSE
Baseline (Average)	0	22.93	914.65	30.24
Linear Regression	0.36	18.00	579.88	24.08
Decision Tree Regressor	-0.08	21.83	989.09	31.45
K-Neighbors Regressor	0.08	21.21	833.15	28.86
Ada Boost Regressor with Decision Tree	0.39	16.81	559.63	23.65
Random Forest Regressor	0.44	16.53	509.27	22.56
Gradient Boosting Regressor	0.44	16.74	514.36	22.67
XGBoost Regressor	0.46	16.11	489.15	22.12
LightGBM Regressor	0.48	15.99	479.14	21.88

Feature Importance



Transformation: Regression → Classification

- Instead of predicting real number, treat it as classification

→ Bin the duration into different periods

- Open question:
 - Deep learning algorithms?

```
if wert >= 75:
    temp['Duration_Label'] = "Very long"
elif wert >= 54:
    temp['Duration_Label'] = "Long"
elif wert >= 36:
    temp['Duration_Label'] = "Medium-long"
elif wert >= 24:
    temp['Duration_Label'] = "Medium"
elif wert >= 14:
    temp['Duration_Label'] = "Medium-short"
else:
    temp['Duration_Label'] = "Short"
```



Next Steps

Next Steps

- Define the target feature
 - → Enrollment count / Enrollment duration
- Enrich country data with information about disease burden
- Possibly enhance information about conditions
- Generate more useful features
- Try out more complex models and apply hyper parameter tuning