FEATURE ENGINEERING & EDA USING EXCEL & MODELING USING PYTHON

Kaggle Titanic Example

BY Kunaal Naik (www.youtube.com/KunaalNaik)

Kunaal Naik

- ✓ Love teaching Data Science
- ✓ YouTube Channel teaching Data Science
- ✓ Marketing Advisor, Data Science - Dell
- ✓ Lifeaholic Evangelist
- ✓ Avid Learner
- ✓ Scuba Diver

(*www.youtube.com/KunaalNaik*)

Agenda

- Why Feature Engineering with Excel?
- Defining the Problem Statement
- Feature Engineering using Excel ***Focus Area
- Build Train and Test Dataset using Excel
- Building Logistic Model using Python
- Evaluation and Submitting Solution to Kaggle Competition
- Warp Up and Q&A

WHY?

Feature Engineering and Pre-processing - Today

Make Pipeline with Column Transformer

```
preprocessor = make_column_transformer(

    (make_pipeline(
    SimpleImputer(strategy = 'median'),
    KBinsDiscretizer(n_bins=3)), numerical_features),

    (make_pipeline(
    SimpleImputer(strategy = 'constant', fill_value = 'missing'),
    OneHotEncoder(categories = 'auto', handle_unknown = 'ignore')), categorical_features),
)
```

Feature Engineering and Pre-processing - Before/Current

Typical Missing Value Treatment Process

```
#Fill Missing numbers with median for Age and Fare
all['Age'] = all['Age'].fillna(value=all['Age'].median())
all['Fare'] = all['Fare'].fillna(value=all['Fare'].median())

#Bin Age
all.loc[ all['Age'] <= 16, 'Age'] = 0
all.loc[(all['Age'] > 16) & (all['Age'] <= 32), 'Age'] = 1
all.loc[(all['Age'] > 32) & (all['Age'] <= 48), 'Age'] = 2
all.loc[(all['Age'] > 48) & (all['Age'] <= 64), 'Age'] = 3
all.loc[ all['Age'] > 64, 'Age'] = 4

#Treat Embarked
all['Embarked'] = all['Embarked'].fillna(value=all['Embarked'].mode()[0])
```

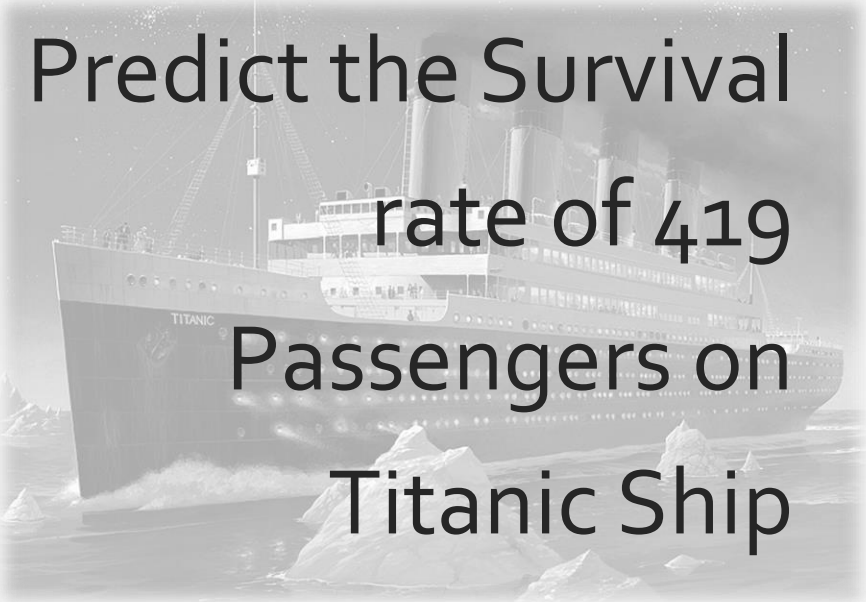
Feature Engineering and Pre-processing Visually!

M	N	O	P	Q	R	S	T	U
1309	1309	1309	1309	1309				
Embarked_1	Fare_1	Age_1	Sex_1	Cabin_1	Family	IsAlone		
S	7.25	22	1	M	2	=IF(R3=1,1,0)		
C	71.2833	38	0	C				
S	7.925	26	0	M	1			
S	53.1	35	0	C	2			
S	8.05	35	1	M	1			
Q	8.4583	26	1	M	1			
S	51.8625	54	1	E	1			
S	21.075	2	1	M	5			
S	11.1333	27	0	M	3			

IF(logical_test, [value_if_true], **[value_if_false]**)

LET REDISCOVER FEATURE ENGINEERING VISUALLY!

Problem Statement



Predict the Survival rate of 419 Passengers on Titanic Ship

891

Passengers (Train)

With Survival information along with their Demographics and other information

419

Passengers (Test)

With Only their Demographics and other information

To help train we have a set of variables. However we will create more using Excel

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Our Goal

To make one submission on Kaggle Titanic Competition and score

0.78 – 0.79

Using Random Forest Random Grid Search

Base Score with Gender Submission is **0.75**
With Logistic Regression it is **0.77**

Feature Engineering using Excel

- Missing Value Treatment
  - Numerical
  - Categorical
- Recoding Variables
- Extract New Features
  - Using existing Numerical Features
  - Using existing Categorical Features
- Creating Dummy Variables (or One Hot Coding)

LET DIVE INTO FEATURE ENGINEERING WITH EXCEL

Content Location

- https://github.com/KunaalNaik/CLL_UPGrad_Mantissa_Webinar
- More Content - GitHub Link

Stay Connected

- www.youtube.com/KunaalNaik
- https://www.linkedin.com/in/kunaal-naik/
- https://github.com/KunaalNaik
- https://www.instagram.com/lifeaholism/
- https://twitter.com/KunaalNaik