Data Preprocessing

Agenda

1. Data Preprocessing
2. Steps involved in model building
3. Problem Discussion
4. Steps to approach modeling for the problem

Questions to discuss

1. What is Data Preprocessing and what are the steps involved in it?
2. Why we need data preprocessing?
3. What are the steps involved in model building?

What is Data Preprocessing?

- Data preprocessing is a collective term used to describe a collection of approaches that help get the data ready for analysis
- It helps us clean and transform the raw data to an efficient format for analysis and modeling
- It is generally very contextual and requires good domain understanding to do this well
- Different datasets need different kinds of preprocessing

What are the steps involved in Data Preprocessing?

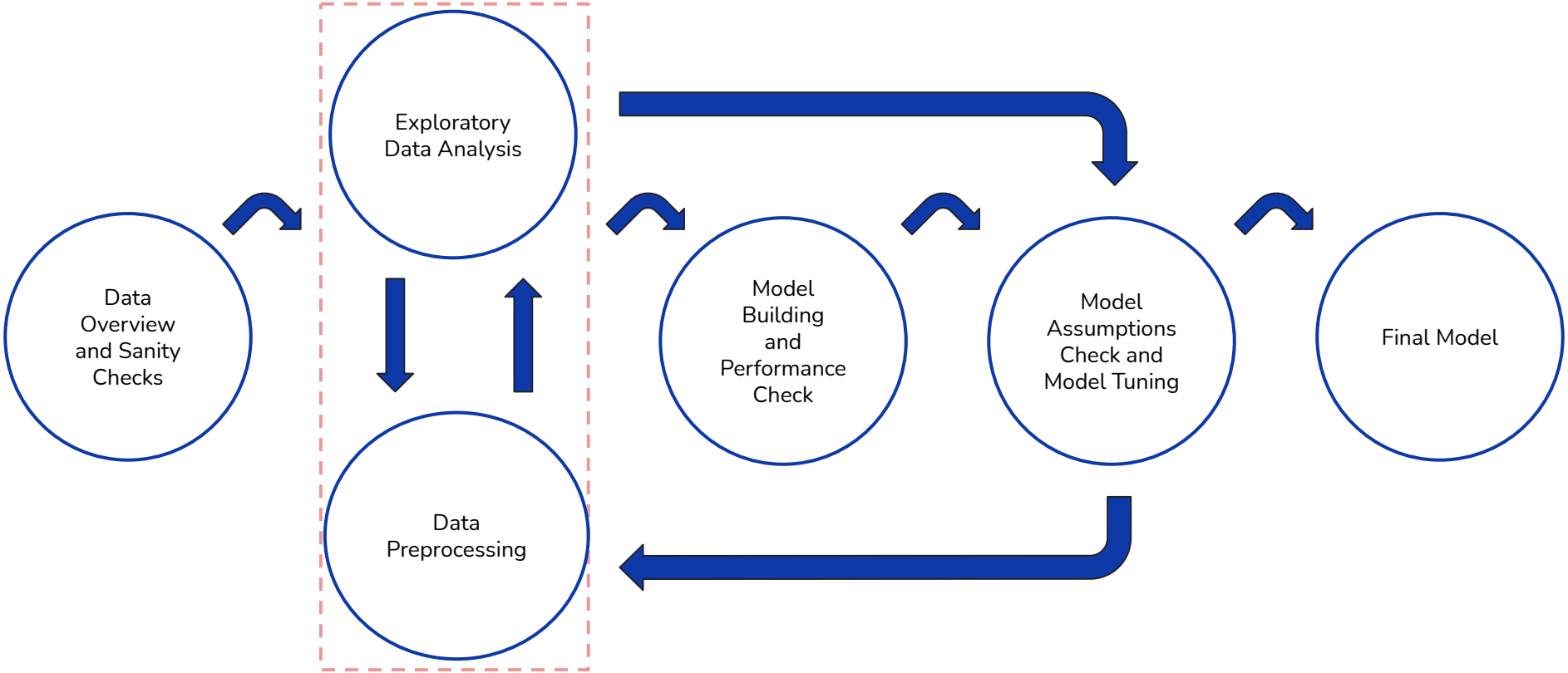
Broadly it is divided into the following steps:

1. Data format checks
    a. Data dimension
    b. Data types
2. Data Consistency
    a. Missing, inconsistent, duplicate values
    b. Outliers
    c. Data distribution and skewness
3. Feature Engineering
    a. Variable transformations
    b. Feature extraction

Why do we need Data Preprocessing?

- Data preprocessing is a crucial step in the cycle of building a model from raw data
- Data preprocessing takes up approximately 60-80% of the time in a modeling project
- We often need to iterate between exploratory data analysis and data preprocessing to obtain the optimal data to get the desired model performance
- Data preprocessing also helps us in case our model does not satisfy any underlying statistical assumptions

What are the steps involved in model building?



Problem - Anime Rating Prediction

- Streamist is a streaming company that streams web series and movies for a worldwide audience.
- Every content on their portal is rated by the viewers, and the portal also provides other information for the content like the number of people who have watched it, the number of people who want to watch it, the number of episodes, duration of an episode, etc.
- They are currently focusing on the anime available in their portal, and want to identify the most important factors involved in rating an anime.
- The objective is to preprocess the raw data, analyze it, and build a linear regression model to predict the ratings of anime.

Anime Data Preview

- Let's take a quick look at a small sample of our data.
- This will help us get some idea of the attributes of the data and also help us understand the degree of cleaning needed before we can build a model.

	title	mediaType	eps	duration	ongoing	startYr	finishYr	sznOfRelease	description	contentWarn	watched	watching	wantWatch	dropped
13764	Spy Penguin (2013): White Christmas	Web	1.0	2.0	False	2013.0	2013.0	NaN	NaN	0	8.0	0	10	0
3782	A Little Snow Fairy Sugar Summer Specials	TV Special	2.0	NaN	False	2003.0	2003.0	NaN	One day, when Saga finds an old princess costu...	0	1056.0	24	576	16
2289	Umineko: When They Cry	TV	26.0	NaN	False	2009.0	2009.0	Summer	In the year 1986, eighteen members of the Ushi...	1	10896.0	1451	8480	1236
5081	Unbreakable Machine-Doll Specials	DVD Special	6.0	5.0	False	2013.0	2014.0	NaN	NaN	1	1957.0	201	756	50
9639	Hanako Oku: Hanabi	TV	1.0	6.0	False	2015.0	2015.0	NaN	NaN	0	46.0	1	54	1

Initial observations from the data

- Two highly textual columns (title and description), can be dropped for modeling
- A lot of the anime have just one episode, indicating they might be movies.
- The duration column has a wide range of values (2 to 90 minutes).
- There is a boolean column (ongoing).
- The sznOfRelease column has a lot of missing values.
- Some other columns too have missing values.

Steps involved in modeling

- Data Overview and Sanity Checks
- Exploratory Data Analysis
- Data Preprocessing
    - Missing Value Treatment
    - Feature engineering (*years_running = finishYr - startYr*)
    - Variable Transformations
    - Outlier Treatment
    - Other preprocessing steps
- Model building and performance check
- Model assumptions check and fixing
- Final model selection

Note: EDA and Data Preprocessing are done iteratively.

Missing Value Treatment

- Drop missing values in the target variable (anime ratings in this case)
- Impute missing values in categorical predictor variables with the mode or a new category ‘missing’
- For numerical variables, we can take the following strategy:
    - Unskewed numerical variable => Impute using the mean
    - Skewed numerical variable => Impute using the median
- However, using the mean/median of the entire data is not always the best approach
    - For example, different media types in which the anime will be published will have different preferable durations.
    - Anime published in TV will likely have episodes with a shorter duration than the ones released as web series or for DVDs.
- As such, it is often a good idea to group the data by such categorical variables and use the grouped mean/median for imputation

Variable Transformations

- Some of the variables may be highly skewed
- Some of the model assumptions may not be satisfied
- Variable transformations are a good way to rectify the above issues
- Common transformations include log, square, square root, exponential, etc.
- Sometimes, it also helps to transform the target variable
    - In case the target variable is highly skewed, using a transform to reduce the skewness often helps to build a better model.

Modeling and Assumptions Check

- Once the data is processed and ready, proceed to modeling
- Choose the metrics to measure the model performance
    - Select the primary metric to monitor as per the business requirements
- Need to split the data into train and test sets
- Build the model on the train set and check the train and test performances
- Check whether the underlying assumptions (if any) of the model are satisfied.
    - Make necessary fixes if one or more of them are violated
- Tune the model performance
    - This can be done by processing the data further and/or testing different algorithms
- Compare and choose the best model

greatlearning
Power Ahead

Happy Learning !