Why data understanding and data preparation steps are important in predictive analytics?

It is important, since without understand the data pattern, trend, history, business problem, context and prepare the data, we can hardly make the correct prediction. First of all, it is important to exploit and validate data. We can do profiling to obtain information about the content and structure of the data. The output of this step is a set of metadata, such as max, min, avg of values. Data validation operations are user-defined error detection functions, which scan the dataset in order to spot some errors. Later, we can clean the data by re-formatting particular attributes and correcting errors in data, such as missing values imputation. Next, we can do the data labelling, where each data point is assigned to a specific category. Finally, we split the data into training, validation, and test datasets to be used during the core machine learning stages to produce the ML model.

Some of the data preparation steps in general are:

- Remove unnecessary data
- Create new features from existing dataset (date and time features, revenue feature, profit features, holiday features, weekday and weekend features)
- Changing the formats
- Cleaning the data
- Removing outliners

In natural language processing, data preparation includes:

- Language detect - recognize which language
- Translation possibly needed
- Removing punctuation, icons, and emojis - depending on context
- Removing stop words
- Remove numeric characters, depending on context
- Lemmatization
- Finding similarities between words
- Stemming
- Tokenization
- Typo handling
- Word2Vec
- Word Embeddings
- If data is from scraping or API - extracting useful text
- OCR/similar for PDF text extraction
- Get sentences out of the text
- All of these steps might be model dependent, make sure your input conforms with the expected inputs
- Continuous Bag of Words

In computer vision, data preparation includes:

- Convert color images to grayscale to reduce computation complexity
- remove the background color from your images
- Standardize/Normalization Image

- Data transformation: Scaling, rotations, flipping and other affine transformations

Explain the difference between descriptive vs predictive modeling with an example.

Both descriptive and predictive models have their use cases in practice, and they can overlape each other and are used at the same time to supplement for each other. Descriptive modelling tells what happens in the past, while predictive modelling tells about the future.

We can define descriptive modeling to exploit the past data to provide an accurate report. As the name may give it away predictive models deals with the predicting the future. It uses data from the past to find patterns to identify any future risks or outcome. In other words, descriptive analytics tell you what happened in the past and predictive analytics will tell you what might happened in the future. Most simple analytics fall under descriptive analytics such as: mean, median, sums, totals, percent, and percent changes when looking ant historical data.

Descriptive modeling is used to describe real-world events and the relationships between factors responsible for them. This is most often used by companies to better target their marketing to the consumer. Customer groups are clustered according to different factors such as demographics, buying habits, interests, and other factors like these. The main aspects of descriptive modeling may include Customer segmentation: Partitions a customer base into groups with various impacts on marketing and service. Value-based segmentation: Identifies and quantifies the value of a customer to the organization. Behavior-based segmentation: Analyzes customer product usage and purchasing patterns. Needs-based segmentation: Identifies ways to capitalize on motives that drive customer behavior. Some other examples are the effectiveness of email or social media campaigns. How webpages perform in terms of clicks, time on page, and conversions. Descriptive modelling is used as a way to looking to the past to predict the future. For example, in marketing, descriptive modelling is used in:

- Customer purchase history
- The effectiveness of email or social media campaigns
- How webpages perform in terms of clicks, time on page, and conversions

Predictive modeling is an application of Predictive analytics, a mathematical process that aims to predict the future events or outcomes based on past behavior or events. As mentioned above it does this by analyzing data to identify patterns that can be forecast what is likely to happen in the future. It all starts with collecting data for analysis, then creating algorithms and statistical models and training them with the subsets of the data and run them against the original dataset to create the predictive model. There are many different modeling methods and algorithms. Some of the most popular ones are, decisions tress, neural networks, time series analysis and many more. Some examples of predictive modelling are:

- Customer and audience segmentation (using cluster modeling)
- New customer acquisition (using identification modeling)
- Lead scoring (using propensity modeling and predictive scoring)
- Content and ad recommendations (using collaborative filtering)

What is the difference between classification and regression problems? Explain which methods are useful in solving classification and regression problems with one example each.

There is some overlapping between classification and regression problems. Predicting price range is the typical multiple classification problem, but predicting price is regression problem, given that the price can be any decimal and there are no limit (as continuous value).

Another classic example is Large-Scale Disk Failure Prediction (PAKDD 2020 Competition and Workshop, AI Ops 2020 February 7 – May 15, 2020).

- The problem of predicting failure can be transformed into a traditional binary classification problem, then we need to predict will the hard disks be damaged or not in the latter 30 days.
- The problem can be transformed into a multiple classification/sorting problem, then we need to predict severity of hard disk damage by learning to rank. If the difference between the hard disk failure time and current time is between 0 days and 30 days. We set the label equal to failure time – current time, which is within [0, 30]. If the difference between the hard disk failure time and current time is larger than 30 days. We set the label equal to 31. If the difference between the hard disk failure time and current time is less than 0 days. We set the label equal to −1. Multiclass classification labelling strategy is much better compared with binary classification. We are able to keep more information compared with binary classification labelling method. The disadvantage of multiclass classification labelling strategy is that it is quite time-consuming if we utilize LightGBM or XGBoost or Catboost for training. Multiclass classification may take several times time compared with binary classification. Besides, it may also lose some information because we set the label of 31 days, 32 days, 33 days ... equal to 31.
- The problem can be transformed into a regression problem, how many days after will the hard disks be damaged since now. We set the difference between the hard disk failure time and current time as our label. Regression methods can utilize information better than multiclass classifi cation. We can utilize more information than multiclass when the difference is larger than 30 days. At the same time, we also relieve the problem of time consuming. Training a regression model is much faster than training a multi class model in this problem. For this reason, our team decided to use regression labelling strategy as our final strategy.

The following table explains the differences between classification and regression.

Regression	Classification
The output variable in regression must be continuous or have a genuine value.	The output variable in classification must be a discrete value.
The regression algorithm's job is to map the continuous output variable (y) to the input value (x) (y).	The classification algorithm's job is to map the discrete output variable to the input value(x) (y).
Continuous data is used with regression algorithms.	With discrete data, classification algorithms are applied.
In regression, identify the best fit line that can more accurately predict the output.	We aim to find the decision boundary in classification to divide the dataset into various classes.
Weather prediction, house price prediction, and other regression problems can be solved using regression algorithms.	Classification algorithms can be used to handle problems like identifying spam emails, speech recognition, and cancer cell identification, among others.
Linear and non-linear regression algorithms are two types of regression algorithms.	Binary Classifier and Multi-class Classifier are two types of classification algorithms.

References:

Materials from the class

PAKDD 2020 Competition and Workshop, AI Ops 2020 February 7 – May 15, 2020

What is the difference between Descriptive and predictive analytics? (emerson.edu)

https://www.educba.com/predictive-analytics-vs-descriptive-analytics/

https://www.geeksforgeeks.org/regression-classification-supervised-machine-learning/ (Sagar Shukla 23 Aug, 2022) https://en.wikipedia.org/wiki/Predictive_analytics

https://se.mathworks.com/discovery/predictive-modeling.html

https://machinelearningmastery.com/data-preparation-is-important/ (Jason Brownlee on June 15, 2020 in Data Preparation)

‘https://www.logility.com/blog/descriptive-predictive-and-prescriptive-analytics-explained/