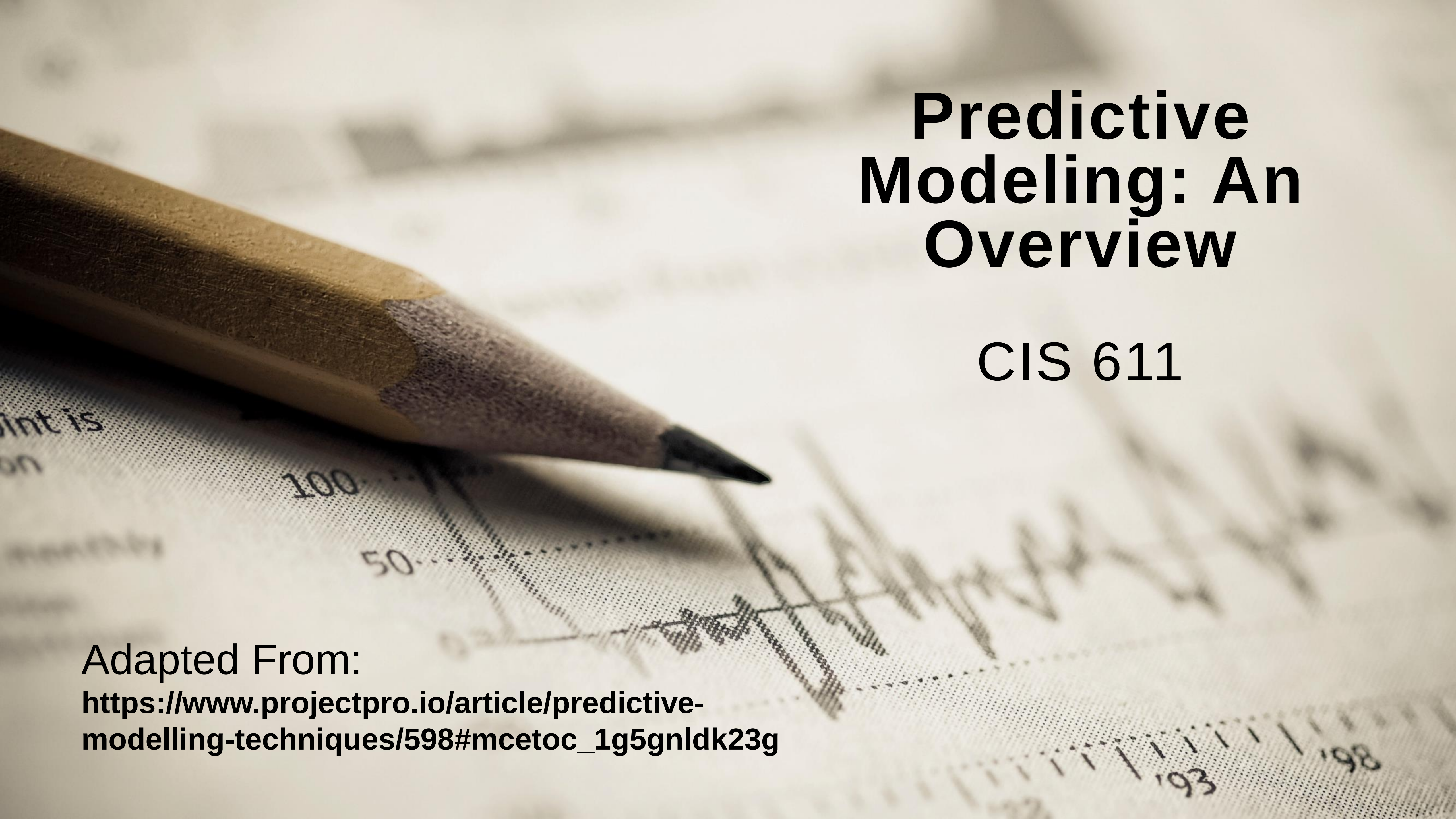

Predictive Modeling: An Overview

CIS 611

Adapted From:

https://www.projectpro.io/article/predictive-modelling-techniques/598#mcetoc_1g5gnldk23g

These slides address, among other questions:

- What is predictive modeling?
- What are the benefits of predictive modeling to businesses?
- What are predictive modeling processes?
- What are the prominent types of predictive modeling techniques (e.g., classification, clustering, forecasting, time series)
- What statistical modeling techniques are used in predictive modeling?

What is Predictive Modeling?

Predictive modeling is a statistical approach that analyzes data patterns to determine future events or outcomes. It's an essential aspect of predictive analytics, a type of data analytics that involves machine learning and data mining approaches to predict activity, behavior, and trends using current and past data.

Banking institutions, for example, may leverage predictive modeling to collect a customer's credit record and other historical data. They might then use this information to calculate a person's credit score and the odds of them making timely credit payments.

How do Predictive Modeling Techniques Benefit Businesses?

Organizations implement predictive analytics using predictive models, which assists them in making better business decisions. Predictive models let companies understand their customer base better, predict future sales prospects, etc. Following are some of the ways in which predictive models benefit various businesses:

- Implement techniques to acquire a competitive advantage,
- Gain a better understanding of the consumer base and their demands,
- Assess and mitigate financial risks,
- Enhance existing products to boost revenue,
- Minimize time and expenses in predicting outcomes,
- Predict external elements that may have an impact on productivity, etc.

Predictive Modeling Examples

Retail- Predictive analytics helps retailers in multiple regions with inventory planning and dynamic pricing, evaluating the performance of promotional campaigns, and deciding which personalized retail offers are best for customers.

By researching consumer behavior and acquiring a better understanding of its customers with the help of predictive models, Staples has achieved a 137 percent return on investment.

Healthcare- The healthcare industry employs predictive analytics and modeling to analyze and forecast future population healthcare needs by leveraging healthcare data. Predictive models in the healthcare industry help identify activities that increase patient satisfaction, resource usage, and budget control. Predictive modeling also enables the healthcare industry to improve financial management to optimize patient outcomes.

The Centre for Addiction and Mental Health (CAMH), Canada's leading mental health teaching center, uses predictive modeling to streamline treatment for ALC patients and maximize bed space.

Predictive Modeling Examples

Banking- The banking industry benefits from predictive analytics by creating a credit risk-aware mindset, managing capital and liquidity, and satisfying regulatory obligations. Predictive analytics models provide more significant detection and protection and better control and compliance. Predictive models allow banks and other financial organizations to tailor each client interaction, reduce customer churn, earn customer trust, and generate remarkable customer experiences.

OTP Bank Romania, part of the OTP Bank Group, implements predictive analytics to govern the quality of loan issuances, yield more precise business and risk forecasts, and meet profit goals for the bank's credit portfolios.

Manufacturing- Manufacturing companies use predictive modeling to forecast maintenance risks and reduce costs on sudden breakdowns. Predictive analytics models help businesses improve their performance and overall equipment efficiency, and also allow companies to enhance product quality and boost consumer experience.

SPG Dry Cooling, a prominent manufacturer of air-cooled condensers, uses predictive modeling to acquire better insights into performance and optimize maintenance, resulting in higher dependability and cost reductions.

Predictive Modeling Overview

- **Predictive Modeling Process**
  - The general steps necessary to acquire and transform data in order to create statistical models and then deploy and monitor those models in practice

- **Application Types of Predictive Models**
  - Categories of statistical models used for prediction aligned to specific industry or business use cases

- **Predictive Modeling Techniques in Machine Learning**
  - Statistical modeling formulas and methods used to accomplish the various types of predictive modeling use cases

Predictive Modeling Process – Part 1

To develop the predictive model, data science experts or analysts generate standard predictive algorithms and statistical models, train them using subsets of the data, and execute them against the entire data set. Let us understand how to build a predictive model using simple and easy-to-understand steps:

- **Data Collection-** The process of data collection is acquiring the information needed for analysis, and it entails obtaining historical data from a reliable source to implement predictive analysis.
- **Data Mining-** You cleanse your data sets through data mining or data cleaning. You delete incorrect data during the data cleansing process, and the data mining process entails removing identical and redundant data from your data collections.
- **Exploratory Data Analysis (EDA)-** Data exploration is essential for the predictive modeling process. You gather critical data and summarize it by recognizing patterns or trends. EDA is the final step in your data preparation phase.

Predictive Modeling Process – Part 2

After curating, cleaning, and exploring data sets for quality and character – the next steps involve building statistical models and establishing protocols evaluate, deploy, and monitor model performance into the future:

- **Predictive Model Development-** You will utilize various techniques to create predictive analytics models based on the patterns you've discovered. Use Python, R, MATLAB, other programming languages, and standard statistical models to test your hypothesis.

- **Model Evaluation-** Validation is a crucial phase in predictive analytics. You run a series of tests to see how effectively your model can predict outcomes. Given the sample data or input sets to evaluate the model's validity, you must assess the model's accuracy.

- **Predictive Model Deployment-** Deployment allows you to test your model in a real-world scenario, which helps in practical decision making and makes it ready for implementation.

- **Model Tracking-** Check the performance of your models constantly to ensure that you are receiving the best future outcomes possible. It involves comparing model predictions to actual data sets.

Application Types of Predictive Models – Part 1

Classification Model

The classification model is one of the most popular predictive analytics models. These models perform categorical analysis on historical data. Various industries adopt classification models because they can retrain these models with current data and as a result, they obtain useful and detailed insights that help them build appropriate solutions. Classification models are customizable and are helpful across industries, including banking and retail.

Clustering Model

The clustering model gathers data and divides it into groups based on common characteristics. Hard clustering facilitates data classification, determining if each data point belongs to a cluster, and soft clustering allocates a probability to each data point.

In some applications, such as marketing, the ability to partition data into distinct datasets depending on specific features is highly beneficial. A clustering model can help businesses plan marketing campaigns for certain groups of customers.

Outliers Model

Unlike the classification and forecast models, the outlier model deals with anomalous data items within a dataset. It works by detecting anomalous data, either on its own or with other categories and numbers. Outlier models are essential in industries like retail and finance, where detecting abnormalities can save businesses millions of dollars. Outlier models can quickly identify anomalies, so predictive analytics models are efficient in fraud detection.

Application Types of Predictive Models – Part 2

Forecast Model

One of the most prominent predictive analytics models is the forecast model. It manages metric value predictions by calculating new data values based on historical data insights. Forecast models also generate numerical values in historical data if none are present. One of the most powerful features of forecast models is that they can manage multiple parameters at a time. As a result, they're one of the most popular predictive models in the market.

Various industries can use a forecast model for different business purposes. For example, a call center can use forecast analytics to predict how many support calls they will receive in a day, or a retail store can forecast inventory for the upcoming holiday sales periods, etc.

Time Series Model

Time series predictive models analyze datasets where the input parameter is time sequences. The time series model develops a numerical value that predicts trends within a specific period by combining multiple data points (from the previous year's data). A Time Series model outperforms traditional ways of calculating a variable's progress because it may forecast for numerous regions or projects at once or focus on a single area or task, depending on the organization's needs.

Time Series predictive models are helpful if organizations need to know how a specific variable changes over time. For example, if a small business owner wishes to track sales over the last four quarters, they will need to use a Time Series model. It can also look at external factors like seasons or periodical variations that could influence future trends.

Predictive Modeling Techniques in Machine Learning – Part 1

There are various statistical techniques that data scientists and analyst use to create appropriate predictive models. Below are several prominent statistical modeling techniques used today:

Linear Regression

One of the simplest machine learning techniques is linear regression. A generalized linear model simulates the relationship between one or more independent factors and the target response (dependent variable). Linear regression is a statistical approach that helps organizations get insights into customer behavior, business operations, and profitability. Regular linear regression can assess trends and generate estimations or forecasts in business.

For example, suppose a company's sales have increased gradually every month for the past several years. In that case, the company might estimate sales in the coming months by linearly analyzing the sales data with monthly sales.

Logistic Regression

Logistic regression is a statistical technique for describing and explaining relationships between binary dependent variables and one or more nominal, interval, or ratio-level independent variables. Logistic regression allows you to predict the unknown values of a discrete target variable based on the known values of other variables.

In marketing, the logistic regression algorithm deals with creating probability models that forecast a customer's likelihood of making a purchase using customer data. Giving marketers a more detailed perspective of customers' choices offers them the knowledge they need to generate more effective and relevant outreach.

Predictive Modeling Techniques in Machine Learning – Part 2

There are various statistical techniques that data scientists and analyst use to create appropriate predictive models. Below are several prominent statistical modeling techniques used today:

Decision Trees

A decision tree is an algorithm that displays the likely outcomes of various actions by graphing structured or unstructured data into a tree-like structure. Decision trees divide different decisions into branches and then list alternative outcomes beneath each one. It examines the training data and chooses the independent variable that separates it into the most diverse logical categories. The popularity of decision trees stems from the fact that they are simple to understand and interpret.

Decision trees also work well with incomplete datasets and are helpful in selecting relevant input variables. Businesses generally leverage decision trees to detect the essential target variable in a dataset. They may also employ them because the model may generate potential outcomes from incomplete datasets.

Gradient Boosted Model

A gradient boosted model employs a series of related decision trees to create rankings. It builds one tree at a time, correcting defects in the first to produce a better second tree. The gradient boosted model resamples the data set multiple times to get results that create a weighted average of the resampled data set. These models allow certain businesses to predict possible search engine results.

The gradient boosted approach often rises as the best technique for overall prediction accuracy in many predictive modeling applications; it may not perform as well on diagnostic metrics for model performance like explainability.

Predictive Modeling Techniques in Machine Learning – Part 3

There are various statistical techniques that data scientists and analyst use to create appropriate predictive models. Below are several prominent statistical modeling techniques used today:

Neural Networks

Neural networks are complex algorithms that can recognize patterns in a given dataset. A neural network is helpful for clustering data and defining categories for various datasets.

There are three layers in a neural network- the input layer transfers data to the hidden layer. As the name suggests, the hidden layer hides the functions that build predictors. The output layer gathers data from such predictors and generates a final, accurate outcome. You can use neural networks with other predictive models like time series or clustering.

Random Forest

A random forest is a vast collection of decision trees, each making its prediction. Random forests can perform both classification and regression. The values of a random vector sampled randomly with the same distribution for all trees in the random forest determine the shape of each tree.

The power of this model comes from the ability to create several trees with various sub-features from the features. Random forest uses the bagging approach, i.e., it generates data subsets from training samples that you can randomly choose with replacement.

Predictive Modeling Techniques in Machine Learning – Part 4

There are various statistical techniques that data scientists and analyst use to create appropriate predictive models. Below are several prominent statistical modeling techniques used today:

ARIMA

ARIMA stands for 'AutoRegressive Integrated Moving Average,' and it's a predictive model based on the assumption that existing values of a time series can alone predict future values. ARIMA models only need previous data from a time series to generalize the forecast. These models manage to boost prediction accuracy while keeping the model simplistic. ARIMA models use differencing to change a non-stationary time series into a stationary one, then use historical data to forecast potential values. These models use auto-correlations and moving averages over residual data errors to generate predictions.

ARIMA models have a wide range of applications in various industries. It is helpful in demand forecasting, such as predicting future demand in the food industry. This is mainly because the model offers managers reliable standards for making supply chain decisions.

Support Vector Machines (SVM)

Support Vector Machines (SVMs) are top-rated in machine learning and data mining. The support vector machine is a data classification technique for predictive analysis that allocates incoming data items to one of several specified groups. In most circumstances, SVM acts as a binary classifier, which means it considers the data has two possible target values. Compared with other classifiers, support vector machines offer reliable, accurate predictions and are less prone to overfitting.

SVM transforms your data using a technique known as the kernel trick and then determines an ideal boundary between the potential outputs based on these alterations. Simply told, it performs some extremely complex data transformations before deciding how to separate your data using the labels or results you choose.