TO EXPLAIN OR TO PREDICT?

The Dual Objectives and Application Differences in Statistical Modeling

Author: Galit Shmueli
Published in: Statistical Science, 2010
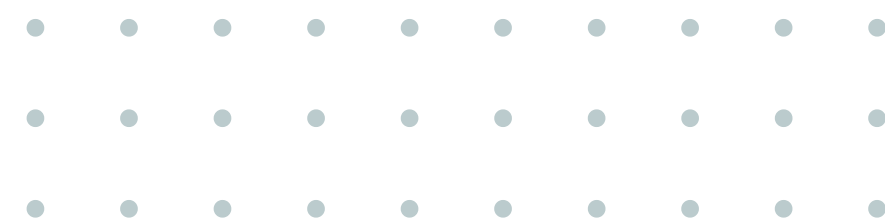
Wanrong Dang
11/04/2024

TABLE OF CONTENT

Introduction

Three Main Applications of Statistical Modeling

- Causal Explanation: Used to uncover causal relationships between variables (e.g., does smoking cause cancer?).

- Prediction: Used to forecast future events or behaviors (e.g., weather forecasting).

- Data Description: Used to summarize and describe the structure of data.

Disciplinary Differences

- Social Sciences: Tend to focus on causal explanation, assuming models with high explanatory power can also predict.

- Technical Fields (e.g., Bioinformatics, NLP): Emphasize prediction to solve practical problems.

Introduction

Research Issue

- Shmueli notes that the confusion between explanatory and predictive modeling can hinder scientific progress.

- Properly distinguishing between these two can improve model effectiveness in research and application.
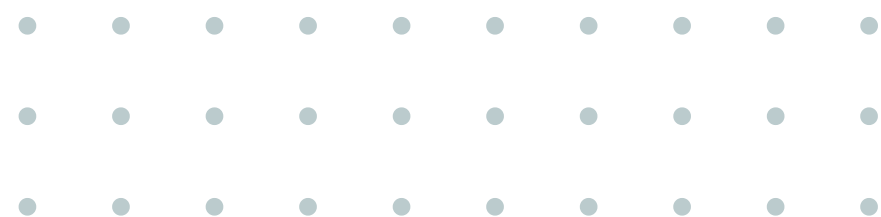
Explanatory Modeling

Definition
Used for causal inference and theory testing. It reveals causal relationships between variables through hypotheses and data analysis.

Application Areas
- Economics: Examines how income affects consumption spending.
- Social Sciences: Studies the impact of education on job opportunities.

Process
- Formulate Causal Hypotheses: Determine causal relationships (e.g., does X cause Y?).
- Select Theoretical Model: Build a model framework based on existing theories.
- Collect and Validate Data: Ensure data quality and relevance.
- Model Testing: Use statistical methods (e.g., regression analysis) to test hypotheses.
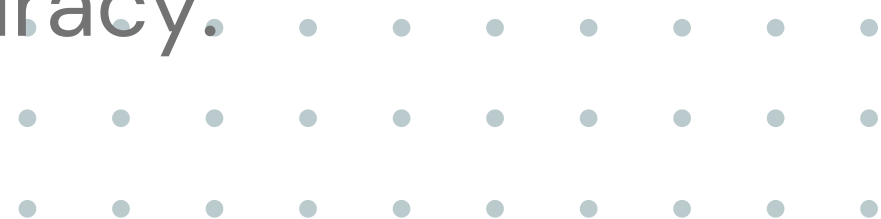
Predictive Modeling

Definition
Used to predict future values of new observations, addressing practical needs.

Application Areas
- Business: Predicts customer purchase behavior and preferences.
- Weather Forecasting: Uses historical weather data to predict future conditions.

Process
- Data Cleaning: Remove noise, ensuring data quality.
- Modeling and Tuning: Use machine learning algorithms (e.g., neural networks, regression analysis) to build the model.
- Model Accuracy Evaluation: Use holdout or cross-validation sets to assess prediction accuracy.
- Model Optimization: Continuously adjust parameters to improve prediction accuracy.
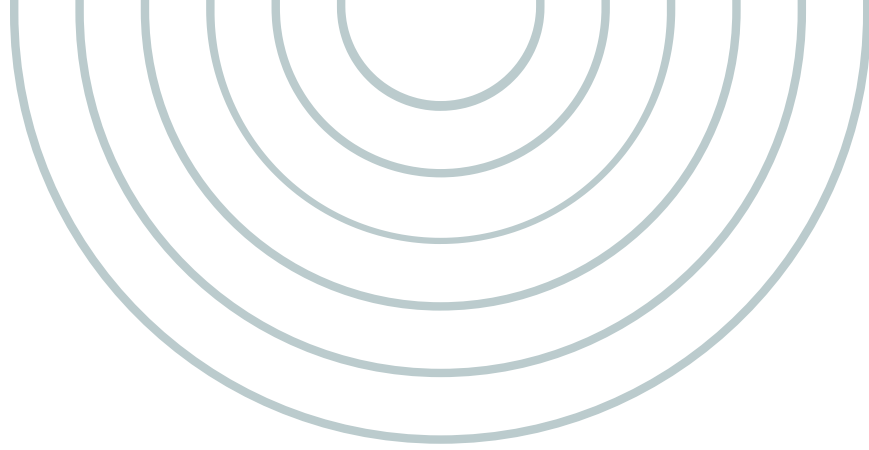
Key Differences

Causality vs. Association

- Explanatory Modeling: Focuses on discovering and validating causal relationships.
- Predictive Modeling: Focuses on the association between variables to forecast future trends.

Retrospective vs. Prospective

- Explanatory Modeling: Retrospective analysis, tests known theories.
- Predictive Modeling: Prospective analysis, forecasts the future.

Theory-Driven vs. Data-Driven

- Explanatory Modeling: Theory-driven, using data to validate hypotheses.
- Predictive Modeling: Data-driven, focusing on maximizing predictive accuracy.

Bias-Variance Tradeoff

- Explanatory Modeling: Leans towards reducing bias to improve theoretical match.
- Predictive Modeling: Allows for bias to reduce variance, ensuring stable predictions.

Model Selection and Evaluation
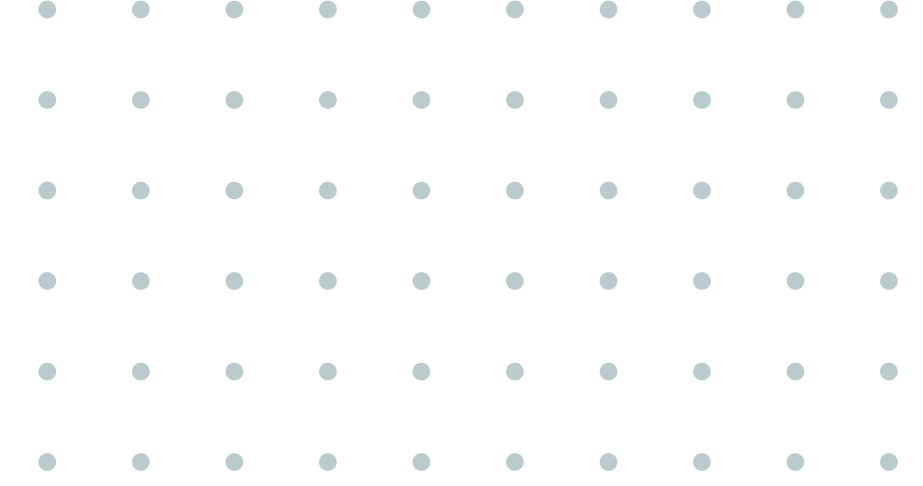
Explanatory Modeling

- **Evaluation Criteria:**
  R^2 value, significance tests, causal inference.
- **Validation Methods:**
  Model fit checks, such as residual analysis and normality testing.
- **Model Selection:**
  Chooses variables based on theoretical relevance and explanatory power.

Predictive Modeling

- **Evaluation Criteria:**
  Prediction accuracy (e.g., MSE, accuracy rate).
- **Validation Methods:**
  Uses holdout or cross-validation to assess generalization, avoiding overfitting.
- **Model Selection:**
  Prioritizes predictive performance over causal explanation, selecting the most accurate models.

Netflix Competition

- Background: Netflix used predictive modeling to improve its movie recommendation system.
- Method: Used machine learning algorithms like matrix factorization and collaborative filtering to enhance predictive accuracy.
- Outcome: The winning team improved accuracy through model combination, earning a million-dollar prize.

eBay Auction Study

- Background: Examined factors influencing final auction prices.
- Method: Used explanatory modeling to determine how seller ratings, auction duration, etc., affect price.
- Outcome: Explanatory modeling helped identify causal influences on auction results.

Real-World Applications

"JUST KEEP MOVING FORWARD."

Impact on Scientific Research

Absence of Predictive Modeling

The lack of predictive modeling can result in inadequate theory validation, hindering new theory development.

Contribution to Research and Practice

- **Academia:** Clearly distinguishing causal explanation and prediction reduces theoretical misuse.
- **Practice:** Influences decision-making, marketing, healthcare, policy, and more.

Complementary Role of Explanatory and Predictive Modeling

- **Explanatory Modeling:** Builds and validates theoretical frameworks.
- **Predictive Modeling:** Provides real-world testing, verifying the practical applicability of theories.

Summary and Recommendations
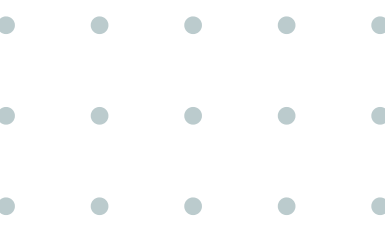
Main Conclusions

- Explanatory and predictive modeling each have unique value.
- Research should select modeling paths based on the specific needs to enhance scientific validity and practicality.

Recommendations

- **Clarify Modeling Objectives:** Decide on explanatory or predictive modeling during study design.
- **Select Appropriate Evaluation Criteria:** Explanatory models focus on explanatory power, predictive models on accuracy.
- **Encourage Predictive Modeling:** Academia should place greater emphasis on predictive modeling to better bridge theory and practice.
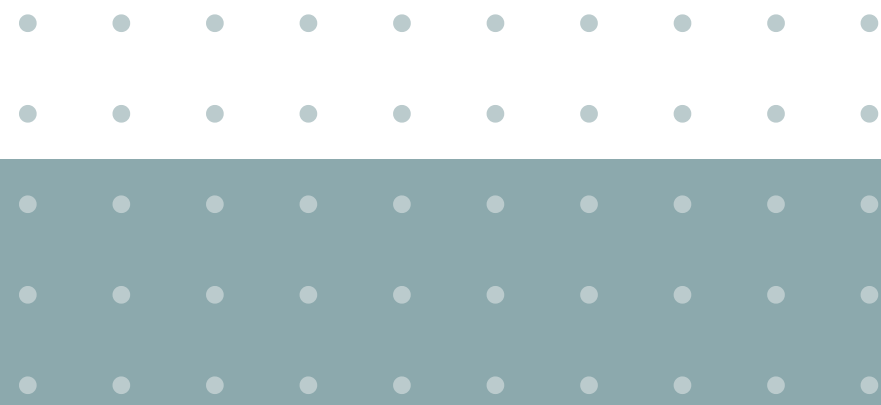
THANK YOU

Do you have any question?