The Real-World Guide to

Data Analytics & Data Mining

From Descriptive to Prescriptive

Prepared by: Dr. Benyawarath “Yaa” Nithithanatchinnapat

Your coffee-stained companion for making sense of data

Data Analytics & Data Mining
Dahlkemper School of Business • Gannon University

Spring 2026

Contents

The Big Picture: Four Types of Analytics....4
Where Each Model Fits on the Spectrum....4
Before You Touch a Model: Data Prep & EDA....5
The Data Reality Check....5
1. Linear & Multiple Regression....6
Sample Questions by Industry....6
Assumptions (And Why They Matter)....6
Key Model Settings & Parameters....6
Model Evaluation: Is It Any Good?....7
2. Logistic Regression....8
Sample Questions by Industry....8
Assumptions....8
Key Model Settings & Parameters....8
Model Evaluation: Classification Metrics....8
3. Decision Trees....10
Sample Questions by Industry....10
The Beautiful Thing About Trees....10
Key Model Settings & Parameters....10
Model Evaluation....11
4. Random Forests & Ensemble Methods....12
Sample Questions by Industry....12
Data Requirements....12
Key Model Settings & Parameters....12
Model Evaluation....12
5. Clustering Methods....14
Sample Questions by Industry....14
K-Means vs. Hierarchical Clustering....14
Data Preparation for Clustering....14
How Many Clusters?....14
Model Evaluation....15
6. Market Basket Analysis....16
Sample Questions by Industry....16
Key Metrics....16
Key Settings....16
Model Evaluation....16
7. Time Series Forecasting....17
Sample Questions by Industry....17

Assumptions 17
Components to Identify 17
Key Model Settings 17
Model Evaluation 18
8. Text Analytics & Sentiment Analysis 19
Sample Questions by Industry 19
Data Preparation for Text 19
Common Gotchas 19
Model Evaluation 19
9. From Prediction to Prescription 21
The Prescriptive Toolkit 21
Connecting Prediction to Action 21
The Golden Rules Across All Models 22
1. Start Simple 22
2. Validate Everything 22
3. Business Understanding > Statistical Perfection 22
4. Document Your Assumptions 22
5. When In Doubt, Visualize 22
6. Check for Bias 22
Quick Reference: Which Model When? 22

The Big Picture: Four Types of Analytics

Before we dig into specific models, let's zoom out. Analytics isn't just one thing — it's a spectrum. Think of it like going to the doctor: first they check your vitals (descriptive), then figure out what's causing the problem (diagnostic), predict whether it'll get worse (predictive), and finally recommend a treatment plan (prescriptive).

Every business analytics project lives somewhere on this spectrum. Knowing where you are helps you pick the right tools and set realistic expectations.

Analytics Type	Core Question	What It Does	Example
Descriptive	What happened?	Summarizes historical data using dashboards, reports, and visualizations	Monthly sales dashboard showing revenue by region
Diagnostic	Why did it happen?	Digs into root causes using drill-downs, correlations, and comparisons	Investigating why Q3 sales dropped 15% in the Midwest
Predictive	What will happen?	Uses statistical models and ML to forecast outcomes or classify cases	Predicting which customers will churn next quarter
Prescriptive	What should we do?	Recommends optimal actions using optimization, simulation, or decision rules	Setting dynamic pricing to maximize revenue across 50 hotel properties

> Pro tip: Most organizations are still working on getting descriptive and diagnostic analytics right. Don't skip these — a predictive model built on data you don't understand is a recipe for disaster.

Where Each Model Fits on the Spectrum

Model / Technique	Primary Type	Can Also Serve
Summary Statistics & Dashboards	Descriptive	Diagnostic (drill-downs)
Correlation & EDA	Diagnostic	Descriptive
Linear / Multiple Regression	Predictive	Diagnostic (coefficient interpretation)
Logistic Regression	Predictive	Diagnostic (odds ratios)
Decision Trees	Predictive	Descriptive (segmentation rules)
Random Forests & Ensembles	Predictive	—
Clustering (K-Means, Hierarchical)	Descriptive / Diagnostic	Predictive (segment scoring)
Market Basket Analysis	Descriptive	Prescriptive (recommendation engines)
Time Series Forecasting	Predictive	Prescriptive (inventory planning)
Text Analytics & Sentiment	Descriptive / Diagnostic	Predictive (sentiment classification)
Optimization & Simulation	Prescriptive	—

Before You Touch a Model: Data Prep & EDA

Here's the thing nobody tells you in textbooks: **80% of your analytics work happens before you build a single model.** This section is your survival guide for the messy reality of real data.

The Data Reality Check

Missing Values: The Disappearing Sock Problem

Missing data is everywhere. The question isn't "do I have missing values?" — it's "what do I do about them?" The answer depends on how much is missing and why.

Severity	How Much?	What to Do	Why This Works
Low	< 5% missing	Delete those rows (if you have plenty of data)	You won't miss what you can't see — minimal impact on results
Moderate	5–20% missing	Replace with median (numbers) or mode (categories)	Preserves your sample size without distorting the distribution much
High	> 20% missing	Create a "missing" category or use advanced imputation	Sometimes what's missing IS the story — e.g., customers who skip income questions may behave differently

> Business reality: In SAS Viya Model Studio, the informative missingness property handles missing values automatically — measure variables get imputed with the observed mean plus an indicator variable, and category variables treat missing as its own level.

Outliers: The Drama Queens of Your Dataset

Not all outliers are bad. A $10 million transaction at a bank might be perfectly legitimate — or it might be fraud. Your job is to figure out which.

Detection Method	How It Works	When to Use
Boxplot / IQR Rule	Flag values beyond 1.5 × IQR from Q1 or Q3	Quick visual scan, works for most continuous variables
3-Sigma Rule	Flag values beyond 3 standard deviations from the mean	When data is roughly normal; common in manufacturing and quality control
Domain Expertise	Ask: "Is this value possible in the real world?"	Always — statistics can spot outliers, but only you can decide if they matter

What to do once you find them: Keep real outliers (they're signal). Fix errors (they're noise). Cap extreme values if they're real but disproportionately influential. And always document your decisions — future you will thank present you.

Quick Data Prep Checklist

- Check variable types: Are numbers stored as text? Are dates formatted correctly?
- Examine distributions: Histograms for continuous, bar charts for categorical
- Look for class imbalance: If only 2% of customers churn, your model will need special handling
- Standardize when needed: Clustering and neural networks require variables on similar scales
- Create a partition: 70% training / 30% validation is the standard split for honest assessment

1. Linear & Multiple Regression

Predicting numbers with a straight face

Answers questions like: "If we increase marketing spend by $10K, how much will sales go up?" Perfect for predicting continuous outcomes — sales, temperatures, prices, wait times. This is your workhorse model and often where you should start.

Sample Questions by Industry

Industry	Sample Predictive Question
Finance	What will this property appraise for, given square footage, location, and condition?
Healthcare	How many days will this patient stay in the hospital based on diagnosis and age?
Retail	How much revenue will this store generate next quarter based on foot traffic and promotions?
Manufacturing	What will the defect rate be if we adjust machine temperature by 5 degrees?
Sports Analytics	How many points will this player score based on minutes played and opponent strength?
Marketing	What's the expected ROI of increasing our digital ad spend by 20%?

Assumptions (And Why They Matter)

Assumption	What It Means	How to Check	What to Do If Violated
Linearity	X and Y have a straight-line relationship	Scatter plot: should look like a cloud around a line, not a banana	Transform variables (log, square root) or try polynomial regression
Independence	Each observation minds its own business	Red flag: time series data or repeated measures from same person	Add time variables or use mixed models
Normality of Residuals	Errors follow a bell curve	Histogram of residuals; Q-Q plot	Often OK with large samples (n > 30); consider transformations
Equal Variance	The spread of residuals stays consistent	Plot residuals vs. predicted — should be random cloud, not a cone	Use weighted least squares or robust standard errors
No Multicollinearity	Predictors aren't carbon copies of each other	Check VIF scores — should be < 5 (some say < 10)	Drop redundant variables, combine them, or use regularization

Key Model Settings & Parameters

Variable Selection Methods

With many potential predictors, trying every combination isn't practical. Here's how to let the algorithm help you find the best subset:

Method	How It Works	Pros	Cons
Forward Selection	Starts with nothing; adds variables one at a time based on lowest p-value below entry cutoff	Fast; good for many predictors	Once added, variables never leave; may miss combinations

Method	How It Works	Pros	Cons
Backward Elimination	Starts with everything; removes weakest variable one at a time	Considers all variables initially	Slow with many predictors; computationally expensive
Stepwise	Combines forward and backward — adds and removes at each step	Most flexible sequential method	Can produce unstable results; overfits small samples
LASSO (L1)	Shrinks coefficients; forces some exactly to zero	Built-in feature selection; handles many predictors well	Can be too aggressive — may drop important variables
Ridge (L2)	Shrinks coefficients toward zero but never exactly to zero	Handles multicollinearity well; stable predictions	Keeps all variables — less interpretable
Elastic Net (L1+L2)	Hybrid of LASSO and Ridge	Best of both worlds; good for correlated predictors	Two tuning parameters to optimize

In SAS Viya Model Studio: Variable selection is handled automatically. For manual control, use the entry significance level (default 0.05) and stay significance level (default 0.05) to control which variables enter and remain in the model.

Model Evaluation: Is It Any Good?

Metric	What It Tells You	Practical Range / Cutoff	Watch Out For
R^2 (R-squared)	Proportion of variance explained by the model	0.3–0.5 is often solid for human behavior; 0.7+ for physical/engineering	Always goes up when you add variables — use Adjusted R^2 instead
Adjusted R^2	R^2 adjusted for number of predictors	Same ranges; penalizes unnecessary variables	Preferred over R^2 for comparing models with different numbers of predictors
RMSE	Average prediction error in original units	Smaller is better; compare to the standard deviation of your target	Sensitive to outliers — a few big misses inflate it
MAE	Average absolute prediction error	Smaller is better; more robust to outliers than RMSE	Doesn't penalize large errors as much — can mask big misses
AIC / SBC (BIC)	Balance of fit and complexity	Lower is better; use to compare candidate models	AIC tends to pick more complex models; SBC penalizes complexity more
VIF	Detects multicollinearity among predictors	< 5 is good; < 10 is acceptable; > 10 is a problem	High VIF doesn't affect prediction accuracy — but makes coefficients unreliable

Reality check: An R^2 of 0.45 predicting human behavior (like customer spending) is often excellent. Don't panic — humans are complicated and messy. An R^2 of 0.45 for predicting something physical (like bridge load capacity) would be terrible.

2. Logistic Regression

When life gives you yes/no questions

Predicts probabilities: Will this customer churn? Will this loan default? Will this patient respond to treatment? Instead of predicting an amount, we're predicting the odds of something happening.

Sample Questions by Industry

Industry	Sample Predictive Question
Finance	Will this loan applicant default within 12 months?
Healthcare	Will this patient be readmitted within 30 days of discharge?
Retail	Will this customer respond to our email campaign?
Insurance	Is this claim likely fraudulent?
Sports Analytics	Will this team make the playoffs based on mid-season stats?
HR / Workforce	Will this employee leave the company within the next year?

Assumptions

Assumption	What It Means	How to Check	What to Do If Violated
Independence	Each row is its own story	Check for repeated measures or clustered data	Use generalized estimating equations or mixed models
Linearity of Logit	The log-odds should be linear with predictors	Bin predictor, calculate log-odds per bin, plot it — should be roughly straight	Transform variables or add polynomial terms
No Perfect Multicollinearity	Variables shouldn't be carbon copies	Correlation matrix < 0.8; check VIF	Drop or combine redundant predictors
Adequate Sample Size	Need enough events (1s) per predictor	Rule of thumb: at least 10–20 events per predictor variable	Reduce predictors, use regularization, or collect more data

Key Model Settings & Parameters

- **Cutoff threshold:** Default is 0.5, but adjust based on business costs. If missing a fraudulent transaction costs $50,000 but a false alarm costs $50, lower the threshold significantly.
- **Variable selection:** Same methods as linear regression (forward, backward, stepwise, LASSO). In SAS Viya, the default entry/stay significance is 0.05.
- **Handling class imbalance:** When events are rare (< 5% of cases), consider oversampling, undersampling, or SMOTE. Also evaluate using AUC rather than accuracy.
- **Regularization:** L1 (LASSO) for sparse models with feature selection; L2 (Ridge) for better prediction when you have many correlated predictors.

Model Evaluation: Classification Metrics

Classification evaluation starts with the **confusion matrix** — a 2×2 table of actual vs. predicted outcomes that gives you true positives, true negatives, false positives, and false negatives.

Metric	What It Tells You	Practical Cutoff	When to Prioritize
Accuracy	Overall % correct predictions	> 70% is decent; but misleading with imbalanced data	Balanced classes only — don't trust it for rare events
Precision	When you predict "yes," how often are you right?	Depends on false-positive cost; > 0.7 is a good start	When false positives are expensive (e.g., spam filter, unnecessary surgery)
Recall (Sensitivity)	Of all actual "yes" cases, how many did you catch?	For critical applications, aim for > 0.8	When false negatives are costly (e.g., cancer screening, fraud detection)
Specificity	Of all actual "no" cases, how many did you correctly identify?	> 0.8 for most applications	When you need to confirm negatives (e.g., ruling out disease)
AUC (C-statistic)	Overall model discrimination ability (area under ROC)	0.5 = coin flip; 0.7–0.8 = acceptable; 0.8–0.9 = excellent; > 0.9 = outstanding	Best single metric for comparing classification models
F1 Score	Harmonic mean of precision and recall	> 0.5 for imbalanced data is often decent	When you need balance between precision and recall
Lift (Top Decile)	How much better is the model vs. random selection in the top 10%?	Lift of 2+ means your model is adding real value	Direct marketing, targeting campaigns — when you can only act on a subset

The golden rule of classification metrics: There's always a tradeoff between precision and recall. You can't maximize both. The business problem determines which one matters more.

3. Decision Trees

The choose-your-own-adventure of analytics

Creates rules anyone can understand: "If income > $50K AND age > 35, then likely to buy premium product." Great for finding natural segments and explaining decisions to non-technical stakeholders. When your VP asks "why did the model predict that?" — a decision tree can actually answer.

Sample Questions by Industry

Industry	Sample Predictive Question
Finance	Which loan applicants are high-risk based on income, credit score, and debt ratio?
Healthcare	Which patients are most likely to miss their follow-up appointments?
Retail	Which customers are most likely to respond to a loyalty program upgrade?
Telecom	What combination of usage patterns signals a customer is about to switch carriers?
Sports Analytics	What player characteristics best predict draft-round selection?
Education	Which students are at risk of not completing their degree?

The Beautiful Thing About Trees

Almost no assumptions! Decision trees don't care about linear relationships, normal distributions, or equal variances. They find patterns in the data regardless of shape.

What Trees DO Care About

Issue	Impact	What to Do
Missing values	Some algorithms (CART) handle them naturally; others need imputation	Create a "missing" category — trees love that. Or use surrogate splits.
Outliers	Trees are naturally resistant. Splits are rank-based, not value-based.	That $1M salary? Tree just sees it as "bigger than $100K." Usually fine.
Class imbalance	Tree may ignore the rare class entirely	Use stratified sampling, adjust prior probabilities, or use cost-sensitive learning

Key Model Settings & Parameters

Parameter	What It Controls	Starting Value	Tuning Guidance
Max Tree Depth	How many levels deep the tree can grow	7 for decision trees; 16 for random forest	5–7 levels for business use (keeps it interpretable). Deeper = more overfit risk.
Splitting Criterion	How the algorithm measures "purity" at each node	Gini (default in SAS Viya)	Gini and Entropy produce similar trees. Chi-square (CHAID) allows multi-way splits.
Min Observations per Leaf	Smallest allowed group at the bottom of the tree	30 (or 0.05% of training data)	Too small = overfit. Increase for noisy data or small samples.
Min Observations to Split	Smallest group that can still be divided	Varies; often 2× the leaf minimum	Higher values produce simpler, more generalizable trees.

Parameter	What It Controls	Starting Value	Tuning Guidance
Pruning	Trims back the tree after growing to optimal complexity	Cost-complexity pruning (default)	Always prune. The maximal tree memorizes; the pruned tree generalizes.

Splitting Criteria Explained

All splitting criteria answer the same question: "Does splitting here make the child nodes purer than the parent?" The difference is how they measure purity.

Criterion	Best For	How It Works
Gini Index	Classification (binary/multi-class)	Measures probability that two random cases from a node are different. Pure node = 0. Maximally impure = approaches 1.
Entropy (Information Gain)	Classification	Measures disorder. Pure node = 0. Maximum entropy increases with more classes. Tends to produce slightly more balanced trees.
Chi-Square (CHAID)	Classification with multi-way splits	Uses statistical significance to determine if a split is meaningful. Allows more than two branches per node.
F-test / Variance Reduction	Regression trees (continuous target)	Measures reduction in variance. Splits that reduce target variance the most win.
Misclassification Rate	Pruning (not recommended for growing)	Simple error rate. Less sensitive than Gini or Entropy for finding good splits; better for evaluating final tree.

Model Evaluation

For classification trees, use the same metrics as logistic regression (accuracy, precision, recall, AUC). For regression trees, use R^2, RMSE, and MAE. Plus:

- **Compare training vs. validation performance.** A big gap means overfitting. Prune the tree or increase minimum leaf size.
- **Check variable importance.** Which inputs appear in the top splits? Do they make business sense?
- **Interpretability test:** Can you explain the tree's rules to a non-technical stakeholder in 2 minutes? If not, simplify.

4. Random Forests & Ensemble Methods

When one tree isn't enough, plant a forest

Takes the decision tree concept and goes wild — builds hundreds of trees, each on a random sample of data and features, then lets them vote. Consistently one of the best performers for almost any prediction task.

Sample Questions by Industry

Industry	Sample Predictive Question
Finance	What is this customer's credit risk score across hundreds of behavioral features?
Healthcare	Which combination of lab results and vitals best predicts sepsis onset?
Retail	Which products will this customer buy next, based on browsing and purchase history?
Energy	What will electricity demand be tomorrow given weather, day-of-week, and economic indicators?
Sports Analytics	How likely is this team to win based on 200+ game statistics?
Cybersecurity	Is this network activity pattern consistent with a security breach?

Data Requirements

The magic of ensembles: **almost no strict assumptions.** Mix of variable types? No problem. Non-linear patterns? They'll find them. Interactions between variables? Automatically detected. Missing values and outliers? More robust than single trees because each tree sees a different sample.

Key Model Settings & Parameters

Parameter	What It Controls	Starting Value	Tuning Guidance
Number of Trees	How many trees to grow in the forest	100–500	More is usually better, but diminishing returns after ~300. Watch computation time.
Max Features per Split	How many features each tree considers at each split	√(total features) for classification; total/3 for regression	Lower = more diversity among trees = less overfitting but potentially more bias.
Max Tree Depth	Maximum depth of individual trees	16 (SAS default) or unlimited	Deeper trees capture more complex patterns but risk overfitting individual trees.
Min Samples per Leaf	Minimum observations in each leaf node	30 for classification; 5–10 for regression	Larger values smooth predictions. Increase for noisy data.
Bagging Fraction	Proportion of data sampled for each tree	~63% (bootstrap default)	Lower values increase diversity; try 0.5–0.8 if overfitting.

Model Evaluation

Same classification or regression metrics as before, plus:

- **Variable importance plots:** Which features drive predictions most? Use these for business storytelling.
- **Out-of-bag (OOB) error:** Each tree is tested on the data it didn't see — gives you a built-in validation estimate without needing a separate holdout.

- **Tradeoff alert:** Random forests are accurate but hard to explain. If your stakeholder asks "why?" use SHAP values or LIME for local interpretability, or build a companion decision tree for explanations.

The Black Box Complaint: "I can't explain this to my boss." Response: Use variable importance to show WHAT matters. Use partial dependence plots to show HOW each variable affects predictions. Use individual SHAP values to explain specific cases.

5. Clustering Methods

Finding the groups you didn't know existed

Discovers natural groupings in your data: customer segments, patient subgroups, product categories. No target variable needed — the algorithm finds patterns on its own. This is **unsupervised learning** — you're not predicting anything, you're exploring.

Sample Questions by Industry

Industry	Sample Descriptive/Diagnostic Question
Retail	What natural customer segments exist based on purchasing behavior?
Healthcare	Are there distinct patient subgroups with different treatment response patterns?
Finance	What types of spending patterns exist among our credit card holders?
Marketing	Which audience segments should we target with different messaging?
Sports Analytics	What playing styles define different types of NBA players?
Manufacturing	Are there distinct patterns of equipment failure across our production lines?

K-Means vs. Hierarchical Clustering

Feature	K-Means	Hierarchical
Pre-specify clusters?	Yes — you choose K upfront	No — produces a tree (dendrogram); you pick where to cut
Cluster shapes	Assumes roughly spherical (circular) clusters	Can handle weird shapes and sizes
Speed	Fast, even on large datasets	Slow for large datasets (computes all pairwise distances)
Stability	Results can change with different starting points	Deterministic — same input always gives same output
Best for	Large datasets; roughly equal-sized groups	Smaller datasets; exploring hierarchical relationships

Data Preparation for Clustering

Clustering is the pickiest technique when it comes to data prep:

- **Always standardize variables.** If income is in thousands and age is in years, the algorithm will focus entirely on income. Use z-scores: (value - mean) / std dev.
- **Missing values are deadly.** You must impute or remove them. Consider if "missing" itself is a meaningful signal.
- **Outliers will form their own clusters.** Remove or cap extreme values, or use robust methods like DBSCAN.
- **Choose your distance metric thoughtfully.** Euclidean for standard analysis. Manhattan for high dimensions. Ward's method for minimizing within-cluster variance.

How Many Clusters?

Method	How It Works	Practical Guidance
Elbow Method	Plot K vs. within-cluster sum of squares; look for the "bend"	Most common approach — but the elbow is often subtle
Silhouette Score	Measures how similar each point is to its cluster vs. other clusters	Ranges from -1 to 1. Average > 0.5 is good; > 0.7 is strong
Business Sense	Can your organization actually act on this many segments?	47 customer segments? Nobody can use that. Aim for 3–8 interpretable groups.
Gap Statistic	Compares clustering structure to a random uniform reference	More rigorous than elbow; computationally intensive

Model Evaluation

Clustering doesn't have a single "right answer" metric like classification does. Evaluation is a mix of statistics and business judgment:

- **Within-cluster cohesion:** Are members of each cluster similar to each other? Lower within-cluster variance is better.
- **Between-cluster separation:** Are the clusters distinct from each other? Greater distance between cluster centers is better.
- **Calinski-Harabasz Index:** Ratio of between-cluster to within-cluster variance. Higher is better.
- **The "So What" test:** Can you name each cluster? Can the marketing team create a campaign for each? If not, revisit.

6. Market Basket Analysis

What goes together like peanut butter and jelly?

Finds products frequently bought together. Powers "Customers also bought" recommendations. Also useful for medical symptoms that co-occur, website pages visited in sequence, or courses students take together.

Sample Questions by Industry

Industry	Sample Question
Retail	Which products are frequently purchased together, and how should we arrange store layouts?
E-commerce	What items should we recommend when a customer adds running shoes to their cart?
Healthcare	Which symptoms or diagnoses tend to co-occur in patient records?
Banking	Which financial products do customers tend to bundle (checking + savings + credit card)?
Education	Which elective courses do students tend to take together?
Streaming	Which shows are commonly watched together by the same subscribers?

Key Metrics

Metric	What It Measures	Formula	Practical Guidance
Support	How often items appear together	P(A and B)	Set minimum support threshold (typically 1–5% depending on catalog size)
Confidence	If A, then how likely B?	P(B | A)	Higher confidence = stronger rule. > 50% is often actionable
Lift	How much more likely together than separate?	Confidence / P(B)	Lift > 1 = positive association. Lift = 1 = independent. Lift < 1 = they avoid each other

Key Settings

- **Minimum support:** Filter out rare item combinations. Start at 1%, increase if you get too many rules.
- **Minimum confidence:** Filter out weak rules. Start at 50%.
- **Maximum items per rule:** Keep rules simple and actionable (2–4 items).
- **Item granularity:** Too many SKUs? Group into categories first. Too few? You won't find interesting patterns.

Model Evaluation

Market basket analysis doesn't have traditional model metrics. Instead:

- **Lift > 1.0** means the association is real (not just two popular items appearing together).
- **Actionability test:** Can you do something with this rule? "Bread + Milk" is obvious. "Diapers + Beer" is the kind of non-obvious insight you're looking for.
- **Redundancy check:** Remove rules that are subsets of stronger rules.

7. Time Series Forecasting

Predicting the future, one timestamp at a time

Forecasting anything with a time stamp: daily sales, monthly inventory, quarterly earnings, hourly website traffic. If it happened over time and you want to predict what's next, this is your tool.

Sample Questions by Industry

Industry	Sample Predictive Question
Retail	What will daily sales look like over the next 30 days, accounting for seasonal patterns?
Finance	What will next quarter's revenue be based on historical trends and economic indicators?
Healthcare	How many ER visits should we staff for next week?
Energy	What will electricity demand be each hour of the day tomorrow?
Supply Chain	How much inventory should we order for the holiday season?
Hospitality	What will hotel occupancy rates be each night next month?

Assumptions

Assumption	What It Means	How to Check	What to Do If Violated
Stationarity	Statistical properties stay constant over time	Plot it! Should not trend up/down. Use Augmented Dickey-Fuller test.	Difference the series (today minus yesterday) or detrend
No Missing Timestamps	Can't skip time periods	Check for gaps in your date sequence	Interpolate, forward-fill, or aggregate to a coarser time unit
Equal Intervals	Measurements evenly spaced	Daily means every day, not most days	Fill gaps or aggregate to consistent intervals

Components to Identify

Component	What It Is	How to Spot It	Example
Trend	Long-term direction (up, down, flat)	Add trendline to time plot	Streaming subscriptions growing 8% annually
Seasonality	Repeating patterns at fixed intervals	Look for regular spikes/dips	Retail sales spike every December
Cyclicality	Longer-term fluctuations (not fixed period)	Multi-year patterns, often economic	Housing market follows economic cycles
Residuals	What's left after removing trend and seasonality	Should look random — no patterns	If patterns remain, your model needs work

Key Model Settings

- **Smoothing parameters:** Exponential smoothing alpha (level), beta (trend), gamma (seasonality) — typically auto-optimized.
- **ARIMA orders:** p (autoregressive lags), d (differencing), q (moving average lags). Start with auto-ARIMA.
- **Forecast horizon:** How far ahead to predict. Accuracy degrades the further out you go.

- **External regressors:** Add holidays, promotions, weather as dummy variables for better accuracy.

Model Evaluation

Metric	What It Tells You	Practical Guidance
MAPE (Mean Absolute % Error)	Average % off from actual values	< 10% is excellent; 10–20% is good; > 20% needs work
RMSE / MAE	Average error in original units	Compare to the scale of your data
Forecast Coverage	Do actual values fall within prediction intervals?	95% prediction interval should contain ~95% of actuals
Visual Check	Does the forecast look reasonable?	Always plot forecasts against actuals — your eyes catch things metrics miss

Missing Values in Time Series

Different rules than other models — you can't just delete rows because that breaks the sequence.

- **Forward fill:** Use last known value (good for stable series).
- **Linear interpolation:** Draw a line between known points (good for smooth trends).
- **Seasonal interpolation:** Use the same period from last year (good for strong seasonal patterns).

8. Text Analytics & Sentiment Analysis

Teaching computers to read between the lines

Extracts insights from text: customer reviews, social media posts, survey responses, support tickets. Can classify sentiment (positive/negative), extract topics, or categorize documents.

Sample Questions by Industry

Industry	Sample Question
Retail	What are customers saying about our new product line in online reviews?
Hospitality	What themes appear most in negative hotel reviews on travel sites?
Healthcare	Can we identify patient concerns from free-text survey responses?
Finance	What is the overall market sentiment from financial news articles and analyst reports?
HR	What topics come up most in employee exit interview comments?
Government	What are constituents' primary concerns in public comment submissions?

Data Preparation for Text

Text data requires a completely different prep pipeline than numeric data:

1. Lowercase everything: "Happy" and "happy" should be the same word.
2. Remove punctuation: Unless emoticons matter for sentiment.
3. Strip stop words: "the," "is," "at" add noise but no signal.
4. Stem or lemmatize: "running," "runs," "ran" all become "run."
5. Convert to numbers: Bag of words (count frequencies), TF-IDF (adjust for common words), or word embeddings (capture meaning).

Common Gotchas

Challenge	Example	Impact	Workaround
Sarcasm	"Oh great, another delay"	Looks positive, actually negative	Use advanced models or manual review; simple word-counting misses this
Negation	"Not bad"	"Not" + "bad" = two negatives?	Create bigrams: treat "not_bad" as a single positive token
Domain-specific language	"Sick" in product reviews vs. healthcare	Opposite meanings in different contexts	Build custom dictionaries per industry
Misspellings & slang	"gr8," "luv," "teh"	Words don't match dictionary	Spell-check preprocessing or character-level models

Model Evaluation

For sentiment classification, use standard classification metrics (accuracy, precision, recall, F1). For topic modeling:

- **Topic coherence:** Do the top words in each topic make sense together? Ask a human to review.
- **Topic distinctiveness:** Are topics clearly different from each other, or do they overlap?
- **Business utility:** Can your team act on the topics identified? "Customer Service Issues" is actionable. "Topic 7" is not.

9. From Prediction to Prescription

Turning "what will happen" into "what should we do"

Prescriptive analytics is the final frontier — once you have a prediction, how do you turn it into an optimal action? This is where analytics becomes truly strategic.

The Prescriptive Toolkit

Approach	What It Does	Example
Decision Rules	Apply if-then logic to model outputs	If churn probability > 0.7, trigger a retention call. If < 0.3, do nothing.
Optimization	Find the best allocation of limited resources	Maximize campaign ROI by allocating budget across channels subject to constraints
Simulation (What-If)	Test scenarios before committing	What happens to revenue if we raise prices 5%? 10%? 15%?
A/B Testing	Experiment with different actions on real customers	Test two retention offers to see which reduces churn more

Connecting Prediction to Action

Every predictive model should answer the question: **"So what do we do about it?"** Here's how to bridge the gap:

- **Set decision thresholds** based on business costs, not just model defaults. A 0.5 cutoff is rarely optimal.
- **Calculate expected value:** Probability × Action Benefit – Action Cost. Only act when expected value is positive.
- **Define the stakeholder:** Who receives the model output? What decision do they make with it? If you can't answer this, the model isn't ready for deployment.
- **Measure impact:** Track whether the action taken based on the model actually improved outcomes. This closes the feedback loop.

> Remember: The best model is the one that gets deployed and creates value. A model with 85% accuracy that makes business sense beats 90% accuracy that nobody understands or acts on.

The Golden Rules Across All Models

1. Start Simple

Don't jump to Random Forests when linear regression might work. Simple models are easier to explain, faster to run, and often just as accurate. Build a baseline model first, then get fancy only if you need to.

2. Validate Everything

- **Split your data:** 70% training / 30% validation is standard. Never touch test data during development.
- **Compare training and validation metrics.** If training accuracy is 95% and validation is 65%, you're overfitting.
- **If results seem too good, they probably are.** Check for data leakage — future information sneaking into your training data.

3. Business Understanding > Statistical Perfection

A model with 85% accuracy that your team trusts and uses beats a 90% model that sits in a notebook. Always ask: "Can someone take action on this?"

4. Document Your Assumptions

Future you (or your colleague) will thank present you for writing down what you assumed and why. Include: which variables were excluded and why, how you handled missing data, what cutoffs you chose and the reasoning.

5. When In Doubt, Visualize

A picture really is worth a thousand p-values. Plot your data before modeling, plot your residuals after modeling, and plot your predictions before deploying.

6. Check for Bias

Models trained on biased data make biased predictions. Check whether your model performs equally across different demographic groups. If it doesn't, investigate before deploying.

Quick Reference: Which Model When?

If You Need To…	Use This Model	Analytics Type
Predict a number (sales, temperature, price)	Linear / Multiple Regression	Predictive
Predict yes/no (churn, fraud, response)	Logistic Regression	Predictive
Explain decisions to executives	Decision Tree	Predictive / Descriptive
Get the best prediction accuracy	Random Forest / Gradient Boosting	Predictive
Find natural customer groups	K-Means / Hierarchical Clustering	Descriptive
Discover what products go together	Market Basket Analysis	Descriptive
Forecast what happens next month	Time Series (ARIMA, ETS)	Predictive
Analyze reviews, comments, or feedback	Text Analytics / NLP	Descriptive / Diagnostic

If You Need To…	Use This Model	Analytics Type
Decide the optimal action to take	Optimization / Simulation	Prescriptive

Remember this above all: Every model is wrong, but some are useful. Your job isn't to find the perfect model — it's to find one that helps make better decisions than guessing. And when someone asks why their model isn't working, 90% of the time the answer is: "Check your data prep." Now go forth and predict things. Coffee's on me when you nail your first deployment.