Applying ML Methods w/ Case Study

HI 743

Ryan Gallagher

Department of Health Informatics and Administration
Zilber College of Public Health
University of Wisconsin - Milwaukee

April 17th, 2025

Overview

The Art of Machine Learning

- Predictive analytics uses machine learning to model relationships between descriptive features and a target outcome.
- Machine learning adopts **inductive learning**: generalizing rules from specific instances.
- Inductive learning does **not guarantee truth**—rules learned from training data may not apply universally.
- A learning algorithm must have an **inductive bias**—a built-in assumption about the patterns to look for.

Human Decisions in Predictive Modeling

- Learning isn't just algorithmic—**human choices shape model outcomes**.
- Critical decisions include:
  - Feature selection and transformation
  - Handling missing data
  - Choice of algorithm and its parameters
  - Evaluation and validation strategy
- These choices affect:
  - Model performance
  - Interpretability
  - Generalizability
- For these reasons, machine learning is both a **science and an art**.

CRISP-DM Phases and Key Questions (1/3)

Business Understanding

- What is the organizational problem being addressed?
- How can a prediction model address this problem?
- Do we have situational fluency?
- Can the organization act on the model's outputs?
- What data is available?

Data Understanding

- What is the prediction subject?
- What domain concepts are relevant?
- What is the target feature?
- Which descriptive features will be used?

CRISP-DM Phases and Key Questions (2/3)

Data Preparation

- Are there data quality issues?
- How will we handle missing values?
- How will we normalize our features?
- What features will we include?

Modeling

- What types of models will we use?
- How will we set the algorithm parameters?
- Have underfitting or overfitting occurred?

CRISP-DM Phases and Key Questions (3/3)

Evaluation

- How accurate is the model?
- Is the model's performance good enough to act on?
- Can the model's decisions be explained?

Deployment

- How will predictions be used in practice?
- How will the model be integrated into business processes?
- How will we monitor and maintain the model?

Key Concepts in Predictive Modeling

- **Target feature:** The outcome we want to predict.
- **Descriptive features:** Information used to make the prediction.
- Predictive modeling assumes a relationship exists between the target and descriptive features.
- We do not need to know the exact form of this relationship—just that one can be learned from data.
- Learning from data means **generalizing from past examples** to make future predictions.

The Role of Data and Patterns in Prediction

- We learn **relationships by observing patterns** in data—this is the core of predictive analytics.
- These patterns allow us to build a model that makes accurate predictions on new, unseen cases.
- Data is used to find regularities that are **statistically consistent**, not universally true.
- Not all patterns are useful; some may be due to chance—**validation** is needed to confirm generalizability.
- Good predictive models **balance pattern detection with caution about overfitting**.

The Bias-Variance Trade-Off

- A model must balance two opposing sources of error:
  - **Bias:** Error from overly simplistic assumptions; leads to underfitting.
  - **Variance:** Error from being too sensitive to the training data; leads to overfitting.
- Highly complex models may capture noise as if it were signal.
- Very simple models may miss important patterns in the data.
- The goal is to find a model that's just complex enough to generalize well.

Generative vs. Discriminative Approaches

- **Generative models** try to model how the data is generated.
  - They capture the full joint distribution of features and outcomes.
  - Examples include Naive Bayes and probabilistic graphical models.
- **Discriminative models** focus on distinguishing between different outcomes.
  - They directly learn the relationship between inputs and the target.
  - Examples include logistic regression and decision trees.
- Discriminative models often achieve higher prediction accuracy, but generative models can offer deeper insight into the data structure.

Predictive vs. Explanatory Modeling

- **Explanatory models** aim to understand the underlying relationship between variables.
  - Emphasis is on interpreting parameters and causal inference.
  - Often assumes a specific form or structure to the relationship.
- **Predictive models** aim to make accurate predictions on new data.
  - Focus is on performance rather than interpretation.
  - Often use flexible or complex algorithms without requiring full understanding of the underlying mechanisms.
- **Key difference:** Explanatory models seek understanding, while predictive models seek accuracy.

Case Study: Predicting Customer Churn

The Churn Problem at Acme Telephonica

- Acme Telephonica (A T) is a national mobile phone provider serving customers across all U.S. states.
- Like many telecom companies, A T faces a persistent challenge: **customer churn**.
- In 2008, A T created a customer retention team that flagged churn risks based on high volumes of support calls.
- These customers were offered special deals to stay—but the strategy wasn't working.
- Despite efforts, churn continued to rise steadily over five years.
- In 2010, A T hired Ross, a **predictive analytics specialist**, to take a new data-driven approach.
- This case study follows Ross's work applying the **CRISP-DM process** to build a churn prediction model.

Understanding the Business Problem

- A T approached Ross not with a technical problem, but a **business challenge**: reducing customer churn.
- Ross's first task was to translate this into a concrete analytics goal.
- A T aimed to reduce churn from **10% to 7.5%**—a target Ross agreed was realistic, based on their data and retention strategy.
- Ross emphasized that model usefulness could not be confirmed until the data had been explored.
- He also needed to assess A T's readiness to adopt analytics-driven decision-making and act on its insights.

Understanding Operations and Data Environment

- Ross met with Kate, leader of the customer retention team, to learn about current churn interventions.
  - Customers making more than 3 support calls in 2 months were flagged as churn risks.
  - These customers were contacted with special offers—typically reduced call rates.
- Ross also consulted Grace, the CTO, to assess data availability and systems.
  - A T had transactional systems for billing and call activity.
  - Historical data and demographics were stored in a central warehouse.
  - Grace was a key ally—both gatekeeper and early champion for the project.
- Ross developed situational fluency by interviewing key departments and learning the business model:
  - Monthly-renewed contracts with bundled call minutes.
  - Overage and peak/off-peak rates shaped customer usage patterns.

Where Predictive Analytics Could Help

Based on his assessment, Ross identified several ways predictive modeling could support A T's goal of reducing churn:

- **Customer Lifetime Value:** Predict long-term value of each customer to retain high-potential (but currently low-spending) users—e.g., college students.
- **Churn Likelihood:** Identify customers most likely to churn soon. Move beyond simple rules (e.g., call volume) to richer, multi-feature machine learning models.
- **Offer Personalization:** Predict which retention offer (e.g., rate reduction, bonus minutes) would most likely convince a specific customer to stay.
- **Network Risk Prediction:** Forecast infrastructure failures using usage and diagnostics data to prevent outages—reducing churn caused by service disruptions.

Selected Focus: Churn Likelihood

After reviewing multiple analytics opportunities, Ross and the A T executive team chose to focus on building a **churn prediction model**. Key reasons included:

- **Data Availability:** Grace (CTO) confirmed that relevant data for churn modeling was accessible.
- **Process Fit:** The model could plug directly into existing workflows—A T already had a retention team taking action.
- **Business Insight:** Executives saw value in uncovering the key drivers of churn—useful beyond just retention.

Other projects were ruled out due to:

- Missing or incomplete data (e.g., no records of retention offer outcomes).
- High operational disruption (e.g., changing to lifetime value-based strategies).
- Weak evidence for impact (e.g., network issues assumed—but not proven—to drive churn).

Beginning the Data Understanding Phase

- With churn prediction selected, Ross began deepening his understanding of A T's data infrastructure.
- Worked closely with Grace (CTO) to explore:
  - What data existed
  - How it was stored and structured
  - Which teams owned or used it
- This process would guide the creation of the **Analytics Base Table (ABT)**.
- Ross iterated between Grace, Kate (retention), and other departments to refine relevant domain concepts and feature ideas.

Key Data Resources Identified at A T

Ross identified five primary data sources essential to building the churn prediction model:

- **Customer Demographics**
  - From the A T data warehouse
  - Basic customer details and long-term identifiers
- **Billing Records**
  - From the billing database
  - Up to 5 years of detailed billing history
- **Call Transaction Logs**
  - 18 months of customer call behavior
  - Key for measuring usage and patterns over time
- **Sales Records**
  - Details of handsets issued to customers
  - Tracked in the sales team's transactional system
- **Retention Intervention Data**
  - Simple logs maintained by the retention team
  - Includes contacts made and outcomes, dating back 12 months

Defining the Prediction Task and Domain Concepts

- The **prediction subject** was defined as a single customer — one row per customer in the ABT.
- The **target feature** was churn:
  - Inactivity for 1 month (no calls or bill payment), or
  - Explicit cancellation or non-renewal.
- **Observation period:** 12 months of customer behavior history.
- **Outcome period:** Predict churn occurring 3 months in advance.
- Ross led workshops to define **domain concepts** influencing churn:
  - Demographics
  - Billing behavior and changes
  - Handset age and type
  - Customer care interactions
  - Call behavior patterns

Descriptive Features in the ABT (1/3)

Feature	Description
BILLAMOUNTCHANGEPCT	Percent change in the customer's bill since last month
CALLMINUTESCHANGEPCT	Percent change in call minutes since last month
AVGBILL	Average monthly bill amount
AVGRECURRINGCHARGE	Average recurring charge per month
AVGDROPPEDCALLS	Average number of dropped calls per month
PEAKRATIOCHANGEPCT	Change in peak-to-off-peak call ratio since last month

Descriptive Features in the ABT (2/3)

Feature	Description
AVGRECEIVEDMINS	Avg. number of received call minutes per month
AVGMINS	Avg. number of call minutes per month
AVGOVERBUNDLEMINS	Avg. minutes over the included bundle
AVGROAMCALLS	Avg. number of roaming calls per month
PEAKOFFPEAKRATIO	Ratio of peak to off-peak calls this month
NEWFREQUENTNUMBERS	Count of new frequently called numbers

Descriptive Features in the ABT (3/3)

Feature	Description
CUSTOMERCARECALLS	Number of customer care calls last month
NUMRETENTIONCALLS	Number of retention team calls received
NUMRETENTIONOFFERS	Number of offers accepted by customer
AGE	Customer's age
CREDITRATING	Customer's credit rating
INCOME	Estimated income level
LIFETIME	Months as an A T customer
OCCUPATION	Occupation category
REGIONTYPE	Type of region (e.g., urban, rural)
HANDSETPRICE	Price of the current handset
HANDSETAGE	Age of current handset
NUMHANDSETS	Number of handsets in the last 3 years
SMARTPHONE	Boolean: is current handset a smartphone?

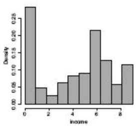
Constructing and Evaluating the ABT

- With Grace's help, Ross integrated all features from Table 9.1 into the **Analytics Base Table (ABT)** using A T's internal tools.
- **Sampling window:** 2008–2013, using churn defined as 1+ month of inactivity (no calls or payments).
- **Active customers** (non-churn): at least 5 calls/week and 6+ months tenure.
- Final ABT: **10,000 customers**, evenly split between churners and non-churners to avoid imbalance.
- Ross created a **data quality report**:
  - **Missingness:**
    - AGE missing 11.47% – imputation possible
    - REGIONTYPE and OCCUPATION had high missingness (74%, 47.8%)
  - **Cardinality issues:**
    - INCOME had only 10 values (banded, not continuous)
    - REGIONTYPE contained inconsistent labels (e.g., `town` vs `t`)
  - Ross cleaned and standardized levels where appropriate

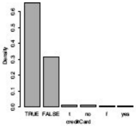
Outlier Analysis and Feature-Target Insights

- Ross identified 4 continuous features with potential **outliers**:
  - `HANDSETPRICE` — minimum value of 0 (e.g., free handsets).
  - `AVGMINS` — max value 6,336.25, far beyond typical usage.
  - `AVGRECEIVEDMINS` — max value 2,006.29, well above average.
  - `AVGOVERBUNDLEMINS` — highly skewed: most values are 0.
- After discussions with Kate and Grace:
  - All outliers were deemed **valid**.
  - `AVGOVERBUNDLEMINS` histogram showed a spike at 0 due to customers not exceeding bundle minutes.
- Ross then explored relationships between features and churn:
  - `REGIONTYPE`: Slightly higher churn in rural areas.
  - `AVGOVERBUNDLEMINS`: Churners used more minutes beyond bundle.
- No single feature showed a strong signal, but multiple **weak patterns** emerged—supporting the use of multivariate modeling.

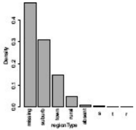
Data Quality Figures



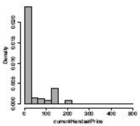
(a) INCOME



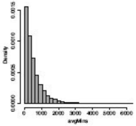
(b) CREDITCARD



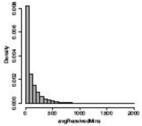
(c) REGIONTYPE



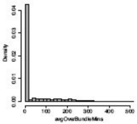
(d) HANDSETPRICE



(e) AVGMINS

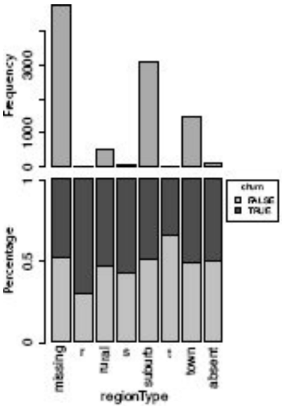


(f) AVGRECEIVEDMINS

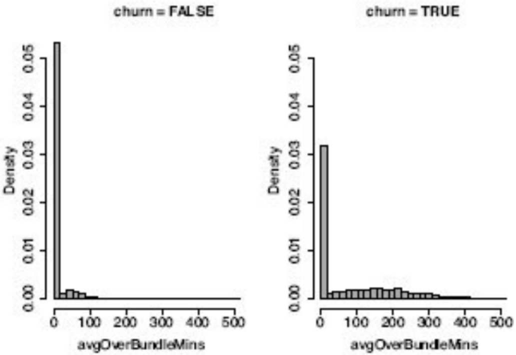


(g) AVGOVERBUNDLEMINS

Data Quality Figures



(a) REGIONTYPE



(b) AVGOVERBUNDLEMINS

Data Quality Report

Feature	Count	% Miss.	Card.	Min.	1^{st} Qrt.	Mean	Median	3^{rd} Qrt.	Max.	Std. Dev.
AGE	10,000	11.47	40	0.00	0.00	30.32	34.00	48.00	98.00	22.16
INCOME	10,000	0.00	10	0.00	0.00	4.30	5.00	7.00	9.00	3.14
NUMHANDSETS	10,000	0.00	19	1.00	1.00	1.81	1.00	2.00	21.00	1.35
HANDSETAGE	10,000	0.00	1,923	52.00	590.00	905.52	887.50	1,198.00	2,679.00	453.75
HANDSETPRICE	10,000	0.00	16	0.00	0.00	35.73	0.00	59.99	499.99	57.07
AVGBILL	10,000	0.00	5,588	0.00	33.33	58.93	49.21	71.76	584.23	43.89
AVGMINS	10,000	0.00	4,461	0.00	150.63	521.17	359.63	709.19	6,336.25	540.44
AVGRECURRINGCHARGE	10,000	0.00	1,380	0.00	30.00	46.24	44.99	59.99	337.98	23.97
AVGOVERBUNDLEMINS	10,000	0.00	2,808	0.00	0.00	40.65	0.00	37.73	513.84	81.12
AVGROAMCALLS	10,000	0.00	850	0.00	0.00	1.19	0.00	0.26	177.99	6.05
CALLMINUTESCHANGEPCT	10,000	0.00	10,000	-16.422	-1.49	0.76	0.50	2.74	19.28	3.86
BILLAMOUNTCHANGEPCT	10,000	0.00	10,000	-31.67	-2.63	2.96	1.96	7.56	42.89	8.51
AVGRECEIVEDMINS	10,000	0.00	7,103	0.00	7.69	115.27	52.54	154.38	2,006.29	169.98
AVGOUTCALLS	10,000	0.00	524	0.00	3.00	25.29	13.33	33.33	610.33	35.66
AVGINCALLS	10,000	0.00	310	0.00	0.00	8.37	2.00	9.00	304.00	17.68
PEAKOFFPEAKRATIO	10,000	0.00	8,307	0.00	0.78	2.22	1.40	2.50	160.00	3.88
PEAKRATIOCHANGEPCT	10,000	0.00	10,000	-41.32	-6.79	-0.05	0.01	6.50	37.78	9.97
AVGDROPPEDCALLS	10,000	0.00	1,479	0.00	0.00	0.50	0.00	0.00	9.89	1.41
LIFETIME	10,000	0.00	56	6.00	11.00	18.84	17.00	24.00	61.00	9.61
CUSTOMERCARECALLS	10,000	0.00	109	0.00	0.00	1.74	0.00	1.33	365.67	5.76
NUMRETENTIONCALLS	10,000	0.00	5	0.00	0.00	0.05	0.00	0.00	4.00	0.23
NUMRETENTIONOFFERS	10,000	0.00	5	0.00	0.00	0.02	0.00	0.00	4.00	0.155
NEWFREQUENTNUMBERS	10,000	0.00	4	0.00	0.00	0.20	0.00	0.00	3.00	0.64

Data Quality Report

(b) Data quality report for categorical features

Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2^{nd} Mode	2^{nd} Mode Freq.	2^{nd} Mode %
OCCUPATION	10,000	74.00	8	professional	1,705	65.58	crafts	274	10.54
REGIONTYPE	10,000	47.80	8	suburb	3,085	59.05	town	1,483	28.39
MARRIAGESTATUS	10,000	0.00	3	unknown	3,920	39.20	yes	3,594	35.94
CHILDREN	10,000	0.00	2	false	7,559	75.59	true	2,441	24.41
SMARTPHONE	10,000	0.00	2	true	9,015	90.15	false	985	9.85
CREDITRATING	10,000	0.00	7	b	3,785	37.85	c	1,713	17.13
HOMEOWNER	10,000	0.00	2	false	6,577	65.77	true	3,423	34.23
CREDITCARD	10,000	0.00	6	true	6,537	65.37	false	3,146	31.46
CHURN	10,000	0.00	2	false	5,000	50.00	true	5,000	50.00

Choosing and Training the First Model

- The model needed to be:
  - **Accurate**
  - **Integrable** into A T's operational workflow
  - **Interpretable**, to give insight into churn behavior
- The ABT contained:
  - A **categorical target feature** (churn vs. no churn)
  - A mix of **continuous and categorical** descriptive features
- **Decision trees** were selected due to:
  - Compatibility with mixed data types
  - Tolerance for missing values and outliers
  - Easy interpretability and business alignment
- Ross trained, tuned, and tested several decision trees:
  - Used entropy-based information gain, binary splits, and no pruning
  - Evaluation metric: **classification accuracy**
  - Initial model achieved **74.87% accuracy** on the hold-out test set

Unpruned, First Fit Tree

Figure 9.4

An unpruned decision tree built for the AT churn prediction problem (shown only to indicate its size and complexity). The excessive complexity and depth of the tree are evidence that overfitting has probably occurred.

Pruned Decision Tree

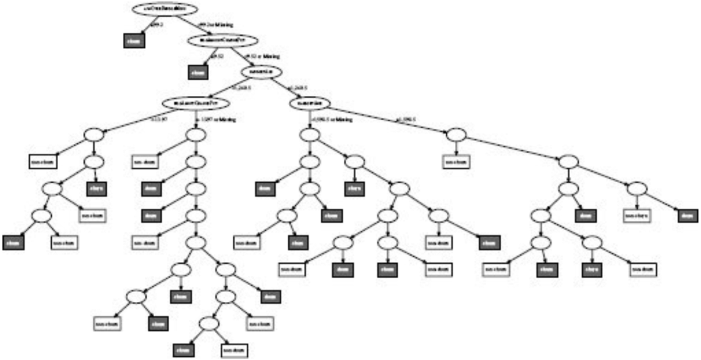


Figure 9.5

A pruned decision tree built for the AT churn prediction problem. Gray leaf nodes indicate a churn prediction, while clear leaf nodes indicate a non-churn prediction. For space reasons, we show only the features tested at the top level nodes.

Confusion Matrix – Stratified Hold-Out Test Set

Using pruning, Ross was able to increase the average class accuracy on the hold-out test set to 79.03%, a significant improvement over the previous model.

Actual / Predicted	**Churn**	**Non-Churn**	**Recall (%)**
Churn	1,058	442	70.53
Non-Churn	152	1,348	89.86

Note: This test set is **stratified**, meaning it preserves the same class balance (churn vs. non-churn) as the training set. This helps ensure fair and representative performance evaluation across both classes.

- The model performs well across both classes.
- **Higher recall** for non-churners than churners - better for predicting non-churners.

Confusion Matrix – Non-Stratified Hold-Out Test Set

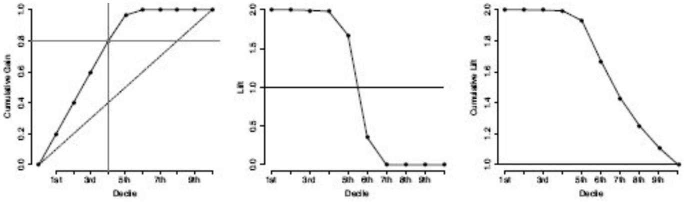
Actual / Predicted	Churn	Non-Churn	Recall (%)
Churn	1,115	458	70.88
Non-Churn	1,439	12,878	89.95

- Results consistent with stratified test set.
- Slight improvement in recall for both classes.
- Indicates good generalization and model stability.

Real-World Evaluation – Non-Stratified Test Set

- The initial **79.03% accuracy** was based on a stratified test set (50:50 churn vs. non-churn).
- In reality, A T's customer base is more like **10% churners, 90% non-churners**.
- To reflect this, Ross evaluated the model on a **non-stratified test set** with the true class distribution.
- Resulting **average class accuracy: 79.28%**, still strong and consistent.
- Ross also used **cumulative gain and lift charts** to evaluate practical performance:
  - The **cumulative gain chart** showed that contacting the top 40% most at-risk customers would capture 80% of likely churners.
  - This demonstrates the model's value in prioritizing retention efforts.

Pruned Decision Tree



(a) Cumulative gain (b) Lift (c) Cumulative lift

Figure 9.6

(a) Cumulative gain, (b) lift, and (c) cumulative lift charts for the predictions made on the large test data sample.

The cumulative gain chart shows that if A T were to call just 40% of their customer base, they would identify approximately 80% of the customers who are likely to churn, which is strong evidence that the model is doing a good job of distinguishing between different customer types.

Presenting the Model to the Business

- With strong performance established, Ross presented the model to broader business teams to build credibility and buy-in.
- For interpretability, he created a **stunted version of the decision tree**, limiting the depth to 5 levels.
  - Full pruned tree was retained for deployment.
  - Shallow trees highlight the most **informative and accessible splits**.
- The simplified tree (next slide) revealed:
  - `AVGOVERBUNDLEMINS`
  - `BILLAMOUNTCHANGEPCT`
  - `HANDSETAGE`

  as key predictors of churn.
- These aligned well with customer behavior: unexpected billing changes, going over call bundles, or aging handsets may drive churn.
- Business stakeholders discussed why features like `CUSTOMERCARECALLS` (used in prior heuristics) were not selected—prompting broader insight.

Pruned Decision Tree

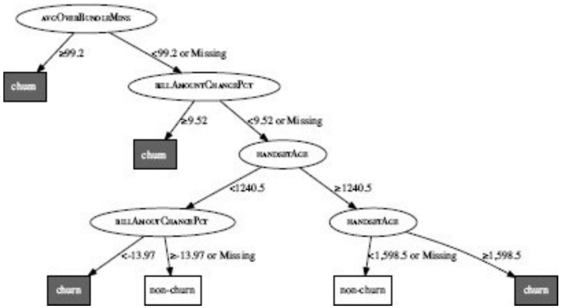


Figure 9.7

A pruned and stunted decision tree built for the Acme Telephonica churn prediction problem.

Deployment and Model Monitoring

- Since A T already had a call list workflow, deploying the decision tree model was relatively straightforward.
- Key deployment tasks:
  - **Revise data pipelines:** Returned to the Data Preparation phase to build robust ETL routines.
  - **Integrate the model:** Code replaced the old rule-based system with the decision tree for generating retention call lists.
- Ross worked with A T's IT team to make the system reliable and repeatable on a monthly basis.
- To ensure long-term effectiveness, Ross implemented a **model monitoring system**:
  - Quarterly reports evaluated prediction accuracy against observed churn (excluding contacted customers).
  - If performance drifted significantly, the model was flagged as stale and retraining was triggered.