Customer Churn Prediction using WEKA

A Comparative Study with Big Data Approaches

Presented by: **Osamah Aldhuwayhi, Hussam Aldossary, Mahmood Jamalullail**

Course: **IS446 - Data Mining**

Instructor: **Dr. Esam Mohammed Othman**



Project IDEA



Business Problem

Telecom companies lose revenue when customers "churn" (switch to competitors) unexpectedly.



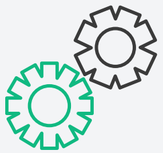
Goal

Understand, analyze, and predict customer behavior to enable proactive retention.



Objective

Use the Telco Customer Churn dataset to build a binary classification model (predict "Yes" churn vs. "No" churn).



Tool

WEKA (Waikato Environment for Knowledge Analysis).

Paper's Methodology

data: Predict churn using massive data (70+ TB, 10M customers) over 9 months.

Platform: Hortonworks Data Platform (HDP) with Apache Spark.



Paper's Methodology



Data Acquisition

Flume & Sqoop to HDFS (Parquet format).

Feature Engineering

- Created >2,000 features
- Two main types:
  - Statistical: averages, counts, max/min per customer
  - Social Network Analysis (SNA): who calls whom, similarity to competitors

Handling imbalance

nly 5% churn (very imbalanced) → tested oversampling / undersampling

Algorithms Used

XGBoost, GBM, Random Forest, Decision Tree

Best Result: XGBoost with SNA features achieved 93.3% AUC.

DATA Preprocessing &Methodology

Step	Action	Tool/Method
1	Load CSV	WEKA Explorer – Open file
2	Handle missing values (TotalCharges)	ReplaceMissingValues filter (replace with mean)
3	Remove customerID	Manual attribute removal
4	Verify attribute types	WEKA automatic detection + manual check
5	Save as ARFF	WEKA – Save button

Final Dataset: 7,043 instances, 19 features, 1 target (Churn).

Algorithm	Parameter	Value	The Reason
Naive Bayes	useSupervisedDiscretization	True	Improves handling of numerical features by converting them into intervals based on class information.
Logistic Regression	maxIts	200	Ensures convergence of the optimization process.
	ridge	1.0E-7	Prevents overfitting by stabilizing coefficient estimates.
J48	minNumObj	5	Reduces overfitting by requiring a minimum number of instances per leaf.
	confidenceFactor	0.15	Produces more aggressive pruning, resulting in a simpler and more generalizable tree.
Random Forest	numIterations	150	Provides a balance between accuracy and computational efficiency.
	maxDepth	12	Limits tree complexity to prevent overfitting.
	seed	1	Ensures reproducibility of results.

Algorithm	Category	Why Selected	Corresponding Algorithm in Paper
Random Forest	Ensemble (Bagging)	Handles non-linear relationships well; robust to overfitting; top performer in many churn studies	Random Forest (used in paper, 87.76% AUC)
J48 (Decision Tree)	Tree-based	Highly interpretable; produces a visual tree; good baseline	Decision Tree (used in paper, 83% AUC)
Logistic Regression	Linear	Simple, fast, interpretable; provides probability scores	Not directly used in paper (paper focused on tree-based methods)
Naive Bayes	Probabilistic	Fast, simple baseline; assumes feature independence	–

Key Results & Screenshots

Algorithm	Accuracy (%)	Kappa	Weighted F-Mesure	ROC Area (AUC)	Precision (Yes)	Recall (Yes)
Naive Bayes	73.08	0.4238	0.746	0.824	0.496	**0.803**
Logistic Regression	**80.22**	**0.4644**	**0.796**	**0.845**	**0.653**	0.545
J48	79.24	0.4323	0.785	0.815	0.635	0.512
Random Forest	79.55	0.4400	0.788	0.833	0.643	0.515

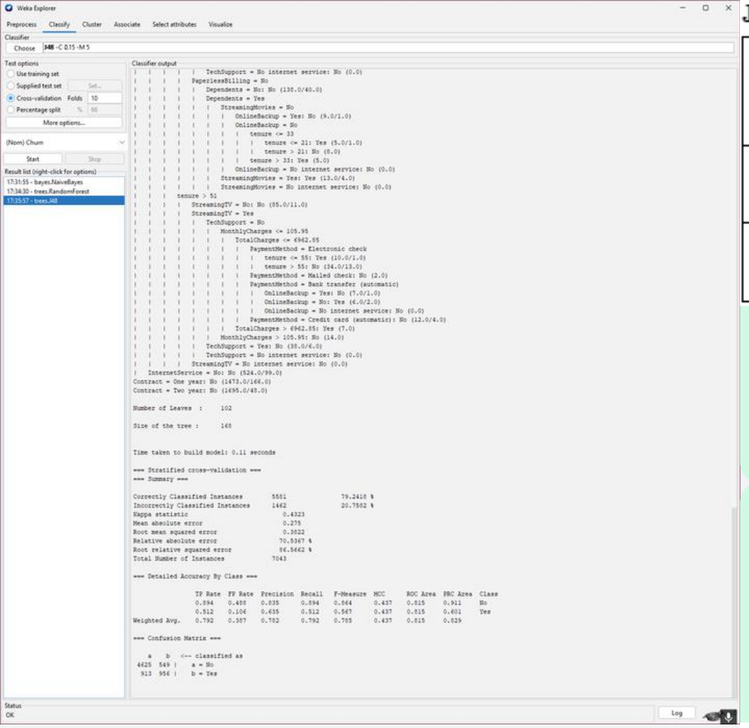
Approach

To ensure reliable performance estimation, 10-fold stratified cross-validation was employed



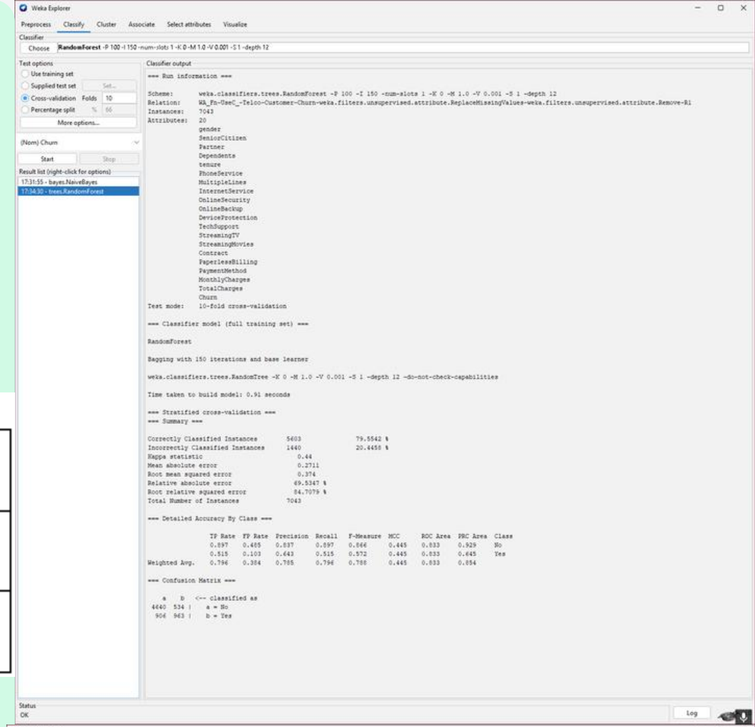
J48 (Decision Tree)

Actual \ Predicted	No	Yes
No	4625	549
Yes	913	956



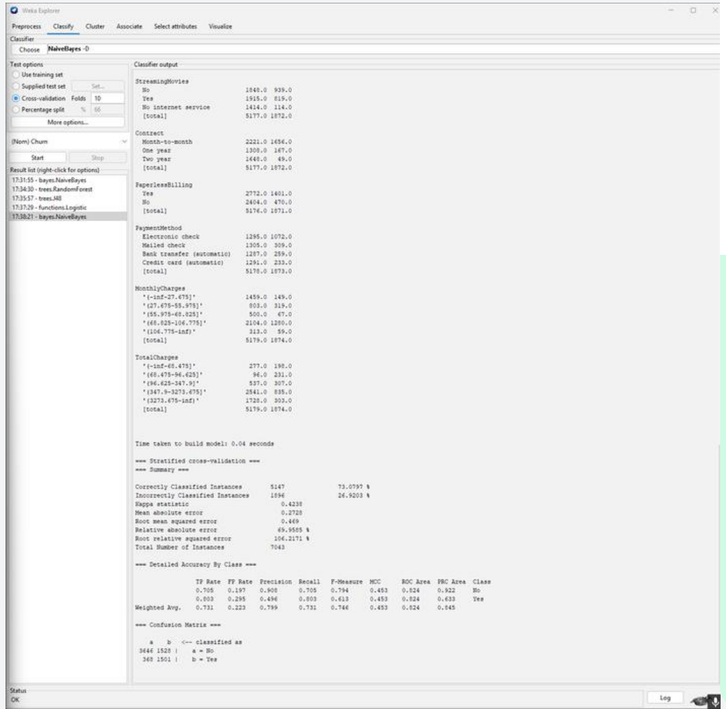
Random Forest

Actual \ Predicted	No	Yes
No	4640	534
Yes	906	963



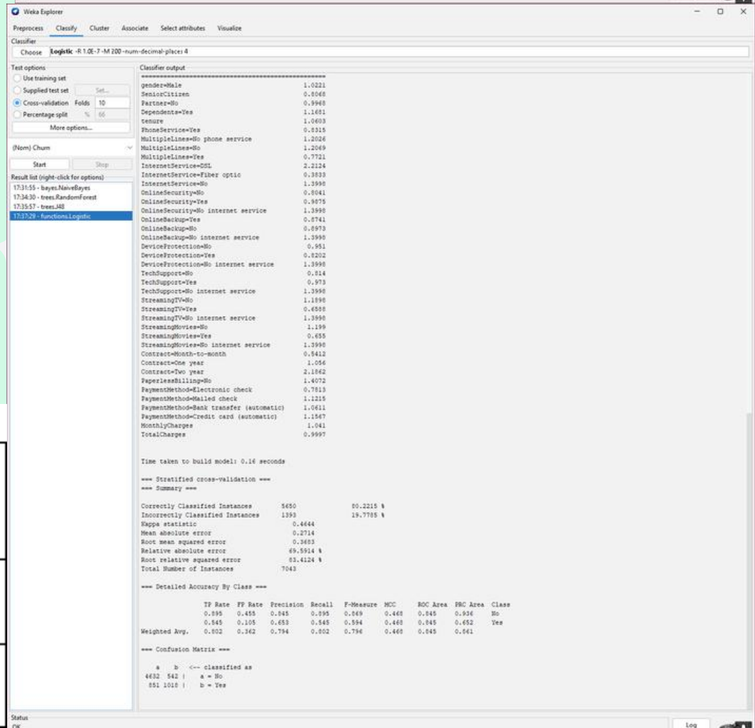
Naive Bayes

Actual \ Predicted	No	Yes
No	3646	1528
Yes	368	1501



Logistic Regression

Actual \ Predicted	No	Yes
No	4632	542
Yes	851	1018



Comparison with Paper's Results

Model (This Study)	Accuracy	AUC	Paper's Equivalent	Paper's AUC
Logistic Regression	80.22%	0.845	Statistical baseline	~84%
Random Forest	79.55%	0.833	Random Forest	87.76%
J48	79.24%	0.815	Decision Tree	83%
Naive Bayes	73.08%	0.824	Not included	–
Best in paper	–	–	XGBoost + SNA	**93.3%**



Discussions

- Why are our results lower than the paper's?
  a. **Data Scale:** Paper used 10M customers + >2,000 features; we used 7K records + 19 features.
  b. **Feature Power:** Paper included Social Network Analysis (SNA) features (e.g., Cosine Similarity, PageRank), which greatly improved prediction.
  c. **Algorithm:** Paper used XGBoost (not in standard WEKA); WEKA has fewer advanced boosting options.
  d. **Class Imbalance:** Paper had 5% churn; we had 26.5% – affects model behavior.
- **Key Trade-off:** High recall (find churners) vs. Low false positives (avoid wasting retention budget).



Challenges Faced



Class Imbalance

- 73.5% Non-Churn vs. 26.5% Churn.
- Models tended to favor majority class. Used stratified CV + multiple metrics (AUC, recall).

Missing Values

- TotalCharges had blanks. Solved with ReplaceMissingValues filter (mean imputation).

Manual Parameter Tuning

- Default parameters gave poor results. Manually tuned (e.g., maxDepth, numIterations).

Software Limitations

- WEKA lacks XGBoost and deep feature engineering capabilities (e.g., SNA).
- Cannot replicate paper's advanced pipeline.

Conclusions



- **Best Model for Accuracy:** Logistic Regression (80.22% accuracy, 0.845 AUC).
- **Best Model for Detecting Churners (Recall):** Naive Bayes (0.803 recall).
- **Practical Finding:** No single model is perfect. Choice depends on business goal:
  - Prevent revenue loss? → Prioritize recall.
  - Minimize retention costs? → Prioritize precision.
- **Comparison Outcome:** Our WEKA models are solid baselines but cannot match the paper's XGBoost + SNA + big data results.

Lessons Learned



1. **Feature Engineering is Crucial:** The paper's massive feature set (especially SNA) dramatically outperformed basic attributes. Good features beat complex models.
2. **AUC is Better than Accuracy:** For imbalanced data, ROC Area (AUC) is a more reliable metric than raw accuracy.
3. **Simple Models Can Work Well:** Logistic Regression performed best on our dataset – don't always need deep learning.
4. **Know Your Tool's Limits:** WEKA is great for learning & small datasets but lacks scalability and advanced algorithms (XGBoost, SNA).
5. **Domain Context Matters:** "High recall" might be more valuable than "high accuracy" in churn prediction (saving a customer is worth some false alarms).

References

- Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1), 1–24. https://doi.org/10.1186/s40537-019-0191-6

- IBM. (2019). Telco customer churn dataset [Data file]. Kaggle. https://www.kaggle.com/datasets/blastchar/telco-customer-churn

- The University of Waikato. (2024). WEKA: Waikato Environment for Knowledge Analysis [Computer software and documentation]. https://www.cs.waikato.ac.nz/ml/weka/



Thank You

Questions are welcome

