Data Science and Engineering (DSE) Record, Volume 7, Issue 1.

Ensemble Collaborative Filtering Technique for Elective Courses Recommender System

Weerinphas Chimnam[1], Dussadee Praserttitipong[2], and Pruet Boonma[3]

[1] Master's Degree Program in Data Science, Chiang Mai University, Chiang Mai, Thailand
weerinphas_chimnam@cmu.ac.th
[2] Department of Computer Science, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand
dussadee.p@cmu.ac.th
[3] Department of Computer Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand
pruet.b@cmu.ac.th

Abstract. This research investigates grade prediction and recommendation for selectable course options under sparse educational data conditions. In this study, elective courses are interpreted as courses for which students have enrollment choices, including Major Elective, Free Elective, and General Education courses where applicable. To address these challenges, the study proposes an ensemble collaborative filtering technique for an elective courses recommender system. The proposed technique integrates collaborative filtering with feature-based regression models and combines historical academic performance, course metadata, and semantic similarity from course descriptions to improve prediction accuracy and coverage. A time-aware evaluation protocol is applied to simulate realistic academic progression and prevent temporal data leakage. Experimental results show that the proposed ensemble models outperform single-model approaches, especially for near-cold-start users, while also maintaining prediction capability for new-item cases. The findings demonstrate that the proposed technique balances accuracy, robustness, and coverage. The system can serve as an additional tool to help students consider selectable course options rather than as a definitive course-selection mechanism.

Keywords: Recommender System, Collaborative Filtering, Educational Data Mining, Ensemble Learning, Grade Prediction, Hybrid Recommender System, Feature-Based Model.

1 Introduction

Higher education institutions increasingly rely on digital academic systems that generate large-scale historical data, including enrollment histories, course descriptions, and performance records. Effectively leveraging such data to support academic planning has become a key challenge in educational data science [1,2]. This study focuses on grade prediction and recommendation for selectable course options, where students may choose among Major Elective, Free Elective, and General Education courses.

CF is widely adopted in educational recommendation for its ability to capture latent student preferences from interaction data [3,4]. However, CF suffers severe performance degradation in sparse scenarios, particularly for new students or newly introduced courses [3,4]. This cold-start limitation is especially critical when curriculum updates create unavoidable interaction gaps. Feature-based models address this by utilizing structured metadata and course descriptions, but they often lack the personalized behavioral signals that CF provides for students with rich histories [6]. This accuracy–coverage trade-off motivates ensemble approaches [7,8].

Despite promising results, many existing studies evaluate ensemble models under simplified settings or with inconsistent data splits that introduce evaluation bias [9]. Near-cold-start users, who have limited but non-zero interaction history, are frequently overlooked despite being a realistic and large subgroup. The contributions of this study are: (1) a leakage-controlled, time-aware evaluation protocol with consistent subgroup masks across all compared models; (2) systematic quantification of ensemble gains specifically for near-cold users across independent temporal evaluation sets; and (3) a reproducible pipeline serving as a methodological template for academic recommender systems. The routing and blending strategies evaluated are established paradigms [7]; the novelty lies in their rigorous evaluation under realistic conditions.

2 Related Work

2.1 Educational Data Mining and Learning Analytics

EDM and LA extract actionable insights from educational data to support learning and institutional decision-making [1,2]. Grade prediction is widely studied as a regression problem using historical performance, demographic attributes, and course characteristics [10,11]. Models must operate under changing curricula and student cohorts, requiring coverage, robustness, and interpretability in addition to accuracy.

2.2 Collaborative Filtering

CF predicts outcomes from historical interaction patterns. Model-based CF using matrix factorization decomposes the interaction matrix into low-dimensional latent factors, where each student u and course i are represented by latent vectors p_u and q_i respectively [3]. The predicted grade is expressed as the sum of a global mean μ, student and item bias terms b_u and b_i that capture systematic deviations in performance, and the inner product of the latent vectors. Bias terms are particularly important in academic data where some students consistently achieve higher grades and certain courses systematically assign lower grades. Recent deep learning extensions such as Neural Collaborative Filtering (NCF) capture nonlinear interactions [16] but require substantially larger datasets and are less suitable for sparse academic settings. The three CF algorithms evaluated in this study, namely KNNBaseline, SVD, and BaselineOnly,

represent neighborhood-based, factorization-based, and bias-only approaches respectively. Despite their effectiveness in warm-start scenarios, CF models face the fundamental cold-start problem: reliable predictions cannot be generated for users or items absent from training data [4,6].

2.3 Feature-Based and Content-Based Models

Feature-based models use engineered attributes including historical GPA, course difficulty, and temporal context [5,10,12]. The prediction function $y_{ui} = f(x_{ui})$ enables full coverage even for new courses. Course descriptions are represented using Term Frequency–Inverse Document Frequency (TF-IDF) weighting, which assigns higher importance to terms that are frequent within a course description but rare across the course corpus [10]. Dimensionality reduction via Truncated SVD then produces dense, low-dimensional embeddings suitable for cosine similarity computation. In addition to structured attributes, semantic similarity features are constructed by computing the cosine similarity between a candidate course embedding and the centroid of a student's historical course embeddings, as well as the centroid of courses associated with the student's academic major [5]. These similarity features enable the model to capture thematic continuity between a student's academic background and the candidate course, providing a content-based signal that is particularly useful when interaction data are sparse or the course has not been seen during training. Despite providing full coverage, feature-based models lack the personalization of CF for warm-start students [6].

2.4 Hybrid and Ensemble Systems

Burke [7] categorizes hybrid recommender systems into three design strategies. Switching selects a single model based on data availability conditions, such as routing to CF when sufficient interaction history exists and falling back to a feature-based model for cold-start cases. Weighted blending combines predictions from multiple models using a tuned weight, allowing complementary signals to be integrated at the prediction level. Feature augmentation uses the output of one model as an additional input to another, enabling deeper cross-model interaction. Prediction-level fusion, as used in switching and blending, offers greater interpretability since each component model remains independently auditable [7].

In academic recommendation, hybrid systems have demonstrated improved robustness for students with sparse enrollment histories, where CF alone cannot generate reliable predictions [7,8]. However, many existing studies apply inconsistent data splits across compared models or allow feature construction to incorporate future information, leading to temporal leakage and overly optimistic performance estimates [9]. Fairness-aware reporting across warm, near-cold, and new-item groups is therefore an essential requirement for rigorous hybrid evaluation, ensuring that observed gains reflect genuine modeling choices rather than evaluation artifacts. The present study addresses these concerns through strict temporal separation, train-only feature

construction, and group-consistent evaluation masks applied uniformly across all compared models.

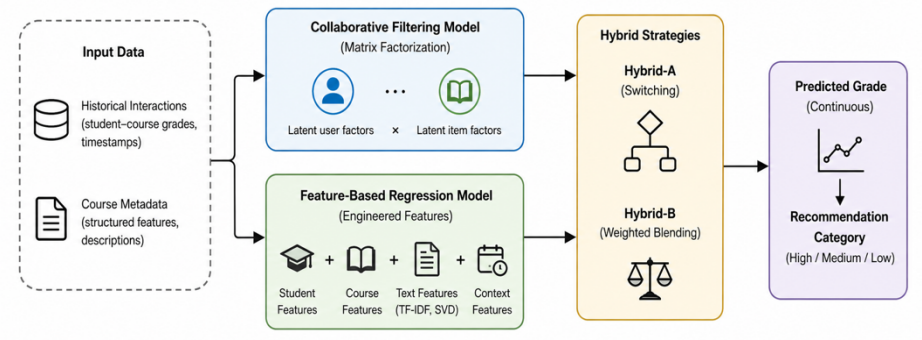
2.5 Research Gaps

Three key gaps motivate this work. First, there is a fundamental accuracy–coverage trade-off: CF achieves high accuracy for warm-start users but fails for new items, while feature-based models provide full coverage but weaker personalization. Second, near-cold users (students with limited but non-zero interaction history) are underexplored despite representing a large and realistic subgroup. Third, evaluation bias from inconsistent splits and temporal leakage is pervasive in existing studies, preventing fair comparison. This work targets all three gaps through ensemble design and a leakage-controlled evaluation framework.

3 Methodology

The framework integrates feature-based regression, CF, and ensemble modeling to achieve both accuracy and full coverage. Fig. 1 illustrates the system overview.

Fig. 1. Overview of the proposed ensemble recommender system.



3.1 Problem Definition and Target Construction

The recommendation problem is defined over a set of students and a set of courses. Historical enrollments are recorded as student–course interaction triples that include the observed grade and the academic time index (year and semester). The prediction objective is to learn a grade estimation function using only information available at prediction time t, explicitly enforcing temporal constraints to simulate realistic deployment. Letter grades are mapped to a GPA scale (A = 4.0, B+ = 3.5, B = 3.0, C+ = 2.5, C = 2.0, D+ = 1.5, D = 1.0, F = 0.0) and normalized to $y_{ui} = r_{ui} / 4.0 \in [0, 1]$. Model performance is evaluated under warm-start, near-cold-start, and new-item conditions.

3.2 Feature-Based Regression

For each pair (u, i), a feature vector x_{ui} is constructed from train-only aggregates to prevent leakage. Features include: (a) course-level mean grade and enrollment count; difficulty proxy $DIFF_i = 1 - \tilde{v}_i$; (b) major-level mean grade; (c) user-level mean grade and enrollment count; (d) major–course mean grade capturing cohort effects; (e) temporal index TermIndex = YEAR×10+SEMESTER and years since entry; (f) cosine similarity between the candidate course TF-IDF/SVD embedding and the centroid of the student's historical embeddings and the student's major centroid.

Four ablation configurations are evaluated: No FE + No Sim, Sim only, FE only, and FE + Sim (full). All aggregates are computed from TRAIN_BASE ∪ SUPPORT only. Ridge regression is selected as the backbone based on minimal-setting performance, computational efficiency, and numerical stability. Ridge minimizes squared prediction error with an l_2-norm penalty on the coefficient vector, controlled by a regularization parameter λ that stabilizes estimation in the presence of correlated and high-dimensional one-hot encoded features. To investigate class imbalance in decision-level evaluation, SMOTE [13] was evaluated as a training-only intervention; however, it degraded classification performance on validation (ACC dropped from 0.714 to 0.616) and was not adopted in the final system.

The feature engineering pipeline is designed around two principles. First, all aggregate statistics must be computed from training data only (TRAIN_BASE ∪ SUPPORT), never incorporating validation or test period records. This constraint ensures that course difficulty proxies, user-level GPA averages, and major–course interaction features do not embed future information, which would artificially inflate prediction performance. Second, categorical variables such as course identifiers and student major codes are processed through one-hot encoding within a scikit-learn ColumnTransformer pipeline [12], with missing values imputed by median (numeric) or a constant 'unknown' token (categorical). Predicted values are clipped to [0, 1] after inference to maintain consistency with the normalized target definition.

The cosine similarity features derived from TF-IDF/SVD course embeddings serve two distinct purposes. The student-history similarity captures thematic continuity between a candidate course and a student's prior coursework, acting as a proxy for curriculum alignment. The major-centroid similarity reflects whether the candidate course falls within the typical course profile of the student's program, providing a coarser but more stable signal particularly useful for new-item courses that share thematic language with established curriculum offerings. Together, these features enable the model to generate predictions for all student–course pairs without relying on any future interaction data.

3.3 Collaborative Filtering

Three CF algorithms are evaluated using the Surprise library [17]. BaselineOnly predicts grades using only global mean and per-student and per-course bias terms, capturing systematic differences in student performance and course grading strictness

without any latent factor learning. SVD extends this with matrix factorization, learning low-dimensional latent vectors for each student and course so that their inner product captures interaction patterns beyond bias. KNNBaseline combines neighborhood aggregation with bias correction, estimating grades by weighting the residuals of the most similar students or courses in the training set. CF is applicable only when both the student and the course appear in the training data, and cannot predict for new-item courses by design.

3.4 Ensemble Strategies

Two strategies follow Burke [7]. Hybrid-A (Routing) applies deterministic model selection: when CF can generate a prediction for the student–course pair, the CF prediction is used; otherwise the feature-based prediction is used as a fallback. This can be written as:

$$\hat{y}_{ui}^{Hybrid-A} = \begin{cases} \hat{y}_{ui}^{CF}, & \text{if CF can predict (u, i)} \\ \hat{y}_{ui}^{feat}, & otherwise \end{cases} \quad (1)$$

This conservative design preserves the strengths of CF under warm-start conditions while ensuring that predictions can still be generated for sparse or new-item cases through the feature-based fallback. Every prediction in Hybrid-A is derived from a single model under a well-defined condition, which makes the routing strategy interpretable and auditable, and therefore suitable for deployment scenarios where explainability is a priority.

Hybrid-B (Blending) uses a weighted combination when CF is applicable:

$$\hat{y}_{ui}^{Hybrid-B} = \begin{cases} \alpha\hat{y}_{ui}^{CF} + (1-\alpha)\,\hat{y}_{ui}^{feat}, & \text{if CF can predict (u, i)} \\ \hat{y}_{ui}^{feat}, & otherwise \end{cases} \quad (2)$$

where $\alpha \in [0, 1]$ is tuned exclusively on validation data. The motivation for blending arises from the observation that near-cold users often have limited but non-zero interaction histories, meaning that CF can provide useful signals that are nonetheless noisy due to weak neighborhood relationships. By combining the CF prediction with the more stable feature-based estimate, Hybrid-B can leverage partial interaction information while reducing the variance introduced by unreliable CF neighborhoods. The blending weight α controls the relative contribution of each component and is selected through a sweep on validation data, ensuring that no information from the test set influences the final configuration. Both strategies ensure full prediction coverage across all student–course pairs, including newly introduced courses for which CF cannot generate predictions by design.

3.5 Time-Aware Evaluation Protocol

Records from years 2559–2563 (B.E.) form TRAIN_BASE; year 2564 Semester 1 is SUPPORT; year 2564 Semesters 2–3 are TEST. VALIDATION is a temporal holdout of year 2563 records from TRAIN_BASE. The final model refits on TRAIN = TRAIN_BASE ∪ SUPPORT ∪ VAL before test evaluation; the test set is accessed only once. Near-cold users are CF-coverable students with support-set interaction count ≤ K = 10 (Q1 of the empirical distribution); a sensitivity analysis over K ∈ {2,3,5,10} confirmed that model performance rankings were consistent across all thresholds. New-item courses arise from newly offered courses absent from training interactions. By construction, fully cold users do not occur in this design since all test-set students appear in the support semester; cold items, however, arise naturally from newly introduced courses in later semesters. These subgroup definitions are used exclusively for analysis and interpretation and do not influence model training or selection in any way.

The time-aware split design has three important properties. First, it reflects realistic deployment conditions: in practice, a recommender system would be trained on all records available at the start of a semester and then asked to predict for the upcoming semester. The SUPPORT set simulates early-semester enrollment data that is available at prediction time but does not yet have final outcomes. Second, all feature aggregates are recomputed from scratch on TRAIN = TRAIN_BASE ∪ SUPPORT before final model refitting, ensuring that course-level statistics, user-level GPA, and major–course interaction features are up to date without using any test-period information. Third, all model selection decisions, including the choice of Ridge over HistGB and Random Forest, the FE+Sim feature configuration, the KNNBaseline CF algorithm, and the $\alpha = 0.65$ blending weight, are conducted entirely on VAL and then frozen before the test set is accessed even once. This 'policy freezing' design, following best practices for leakage-controlled evaluation [9], ensures that reported test-set results reflect true out-of-sample generalization rather than overfitting to the evaluation period.

3.6 Evaluation Metrics

Regression performance is quantified by RMSE and MAE on the normalized [0, 1] scale. RMSE weights larger errors more heavily through its squared formulation, while MAE provides a threshold-independent measure of typical prediction magnitude. For decision-level evaluation, continuous predictions are mapped to three recommendation categories (Recommend, Neutral, and Not Recommend), using thresholds t_neutral = 0.52 (≈2.08 GPA) and t_recommend = 0.73 (≈2.92 GPA), tuned on validation and frozen for test. Classification performance is assessed by accuracy (ACC), balanced accuracy (BAL_ACC), and macro and weighted F1-scores [18]. Balanced accuracy computes the average recall across all recommendation classes, giving equal weight to each class regardless of frequency, which ensures that the minority "not recommend" category is not overshadowed by the majority class. Both macro and weighted F1 are reported to capture aggregate and minority-class behavior simultaneously. Performance is reported for Warm, NearCold, SeenItem, and NewItem subgroups.

4 Data

4.1 Academic Records

The dataset comprises anonymized records from CMU's academic information system across years 2559–2564 (B.E.). Three subsets serve distinct roles. The FINISH table contains completed enrollments with final grades and is the primary source for training, validation, and test targets. The NOW table (ongoing enrollments, no final grades) enriches course-level aggregate statistics in feature engineering only; it is excluded from all regression targets to prevent label leakage. The OUT table (withdrawn/retired students) is excluded entirely as outcomes are incomplete and unreliable. All identifiers are replaced with anonymized keys using secure hashing; records with missing course identifiers or grades are removed. Both FINISH and NOW datasets share the same column schema, including student identifiers, course information, academic term, and curriculum metadata, which allows seamless integration during preprocessing. The structural consistency ensures that NOW records can be incorporated into the feature engineering pipeline without introducing inconsistencies in the aggregate statistics derived from FINISH. To protect student privacy further, the anonymization procedure uses a one-way hashing function that preserves identity consistency across courses and semesters while preventing re-identification of individuals.

4.2 Course Descriptions and Metadata

Structured metadata include course identifiers, credit values, prerequisite relationships, and workload distributions. Bilingual (Thai and English) textual descriptions capture objectives, content scope, and learning outcomes. Descriptions are processed with TF-IDF vectorization and Truncated SVD, without applying stemming or aggressive token filtering, to produce fixed-length semantic embeddings supporting feature-based prediction for new-item courses.

4.3 Dataset Statistics

TRAIN_BASE contains 77,033 records; VAL 13,698; SUPPORT 14,498; TEST 13,205, giving a total of 105,229 records across all splits. The normalized grade distribution has mean ≈ 0.74 and SD ≈ 0.22, corresponding to approximately 2.96 GPA, which reflects a right-skewed distribution where most students receive passing grades. The majority class in the three-category decision mapping is therefore the Recommend category, which has implications for class imbalance in classification evaluation. A global-mean trivial predictor, which assigns the training-set mean normalized grade to every student–course pair regardless of individual or course-specific features, achieves RMSE ≈ 0.21–0.24 on the [0, 1] scale; all learned models must substantially outperform this bound to demonstrate that they capture meaningful predictive signal beyond the overall grade distribution.

Near-cold users comprise 38.2% of the test set (5,039/13,205); new-item courses account for 6.3% (834/13,205). These proportions confirm that sparse-data scenarios are not rare edge cases but constitute a substantial portion of realistic deployment conditions. CF coverage on the test set is 91.87%, meaning that approximately 8.1% of test instances involve courses not seen during training and therefore cannot be handled by interaction-based models alone. The distribution of interaction counts across students is uneven: a small proportion of students have extensive enrollment histories spanning multiple semesters, while a substantial portion have fewer than ten completed courses recorded in the support set, placing them in the near-cold category. This imbalance motivates the use of ensemble models that can leverage partial interaction signals for near-cold users rather than discarding them entirely or treating them identically to fully warm users. Grade distributions also vary across academic programs and course levels, with some major-specific courses exhibiting notably different average grades compared to general education offerings, which is reflected in the major-level and major–course aggregate features included in the feature engineering pipeline. EDA was conducted solely for sanity checking of data quality and distribution; no insights derived from EDA were used to influence feature selection, model training, or hyperparameter tuning, in accordance with the leakage-controlled experimental design.

5 Experimental Results

5.1 Dataset and CF Coverage

Before analyzing predictive performance, it is necessary to verify that the evaluation setup accurately reflects realistic academic conditions and that CF is applicable to a substantial portion of the dataset. Under the time-aware evaluation protocol, CF coverage is expected to be high for users and items observed in the training data, while decreasing for newly introduced courses or previously unseen interactions in later semesters. This characteristic directly motivates the use of ensemble models alongside single-model baselines, since a system that relies solely on CF would fail to generate predictions for approximately 8% of test instances. The reduction in CF coverage from 98.26% on VAL to 91.87% on TEST is therefore not a limitation of the experimental design but an intended consequence of evaluating under realistic conditions where curriculum evolution is unavoidable.

Two additional observations from Table 1 are worth noting before proceeding to model comparisons. First, near-cold users constitute 38.2% of the test set, confirming that sparse-data scenarios are not rare edge cases but a substantial portion of deployment conditions. Second, the mean normalized grade of 0.74 implies that a global-mean trivial predictor achieves RMSE $\approx$ 0.21–0.24 on the [0, 1] scale; this value serves as a natural lower-performance bound, and any learned model must substantially outperform it to demonstrate that meaningful predictive signal has been captured beyond the overall grade distribution. Table 1 summarizes the full dataset split sizes,

CF coverage statistics, and subgroup proportions used throughout the experimental evaluation.

Table 1. Dataset split sizes, CF coverage, and subgroup statistics.

Metric	Split	Value
Training base (TRAIN_BASE)	Train	77,033
Validation set (VAL)	VAL	13,698
Support set (SUPPORT)	SUPPORT	14,498
Test set (TEST)	TEST	13,205
CF coverage	VAL	98.26%
CF coverage (final model)	TEST	91.87%
Near-cold users (K = 10)	TEST	5,039 / 13,205 (38.2%)
New-item courses	TEST	834 / 13,205 (6.3%)
Mean normalized grade (μ)	Train	0.74 (SD ≈ 0.22)

CF coverage is 98.26% on VAL and 91.87% on TEST; the reduction reflects newly offered courses in later semesters. Near-cold users (38.2%) and new-item courses (6.3%) represent realistic deployment scenarios. All learned models substantially outperform the global-mean trivial predictor (RMSE ≈ 0.21–0.24), confirming meaningful predictive signal.

5.2 Regressor Selection (No FE + No Sim)

To ensure a fair comparison of regression models without the influence of feature engineering, candidate regressors are evaluated under the simplest feature configuration, which excludes both engineered aggregate features and semantic similarity features. This minimal setting isolates the effect of the regression model itself, so that any observed differences in performance can be attributed solely to model characteristics rather than to the quality of input representations. Three regression approaches are evaluated: Ridge regression as a regularized linear model, Histogram-based Gradient Boosting (HistGB) as a tree-based ensemble method [15], and Random Forest as a bagging-based ensemble [14]. The selection of a regression model is not determined solely by predictive accuracy, but also by considerations of computational efficiency, reproducibility, and suitability for deployment under time-aware evaluation constraints that require repeated model fitting across temporal partitions.

Ridge regression minimizes squared prediction error with an l_2-norm penalty on the coefficient vector, controlled by a regularization parameter that stabilizes estimation in the presence of correlated and high-dimensional one-hot encoded features. This regularization property is particularly relevant in this study, where categorical variables such as course identifiers and student major codes are one-hot encoded and combined with numerical aggregate features, leading to potentially high-dimensional feature spaces. HistGB improves on standard gradient boosting by discretizing continuous features into bins, reducing memory usage and accelerating training, while Random Forest constructs multiple decision trees using bootstrapped samples and random

feature subsets, averaging predictions to reduce variance. Both nonlinear ensemble methods are included to assess whether capturing complex feature interactions provides additional benefit over the simpler linear model in this sparse academic data setting. The comparison is conducted on the validation set using RMSE and MAE as primary metrics, with computational cost reported alongside predictive accuracy to reflect practical deployment constraints.

Table 2. Feature regressor comparison under minimal feature setting on VAL.

Model	RMSE	MAE	Fit Time (s)	Pred Time (s)
Ridge (selected)	0.1871	0.1440	0.51	0.03
HistGB	0.2034	0.1636	45.15	0.86
Random Forest	0.2081	0.1709	50.73	0.32

Ridge achieves the lowest RMSE (0.1871) with 0.51 s fit time, substantially lower than HistGB (45.15 s) and Random Forest (50.73 s). Nonlinear models do not improve accuracy under sparse feature representations [14,15], confirming Ridge as the appropriate backbone.

5.3 Feature Ablation

After selecting Ridge as the regression backbone, the contribution of different feature components is evaluated through a structured ablation study. Four configurations are compared to isolate the effects of engineered academic aggregates and semantic similarity features independently, as well as in combination. The first configuration excludes both feature types and serves as a minimal baseline reflecting model performance with raw input only. The second incorporates similarity features alone, derived from TF-IDF course description embeddings, to assess their standalone contribution. The third includes only engineered aggregate features constructed from structured academic history, evaluating their predictive value without any textual representation. The fourth combines both feature types and represents the full proposed configuration. By comparing these four settings under identical modeling conditions, the ablation study enables a controlled and interpretable assessment of which components drive predictive performance and how they interact.

Table 3. Feature ablation results using Ridge regression on VAL.

Feature Setting	RMSE	MAE	Fit Time (s)	Pred Time (s)
FE + Sim (selected)	0.1761	0.1360	9.45	0.16
FE only	0.1762	0.1361	0.77	0.07
Sim only	0.1871	0.1439	6.95	0.06
No FE, No Sim (baseline)	0.1871	0.1440	0.23	0.02

FE + Sim achieves the best performance (RMSE = 0.1761), a 5.9% improvement over No FE + No Sim. Structured academic features drive most of the gain (FE only: 0.1762), while semantic similarity features add a small but consistent increment and, importantly, enable coverage for new-item courses. New-item

validation performance (RMSE = 0.1782, N = 238) is comparable to seen-item (0.1761), suggesting that similarity features provide reasonable coverage for new items under controlled validation conditions, though generalization does not extend to test set performance.

5.4 Collaborative Filtering Model Selection

CF models are trained using the same interaction data defined for feature-based models, specifically TRAIN_BASE ∪ SUPPORT, and evaluated exclusively on coverable rows where both the student and the course appear in the training data. This restriction ensures that CF is assessed within its valid operating conditions, where sufficient interaction data are available for model inference. The objective of this evaluation is to examine the predictive performance of different CF algorithms under warm-start scenarios and to establish a strong interaction-based baseline for subsequent ensemble integration. Three representative CF approaches are compared: BaselineOnly, which captures only global mean and bias effects without latent factor learning; SVD, which extends this with matrix factorization to model student–course interaction patterns; and KNNBaseline, which incorporates neighborhood similarity with bias correction. All models are implemented using the Surprise library [17] with consistent handling of train and test splits.

Table 4. CF model comparison on VAL (coverable rows only).

CF Model	RMSE	MAE	Fit (s)	Pred (s)	Coverage
KNNBaseline (selected)	0.1666	0.1240	0.13	0.24	98.26%
BaselineOnly	0.1681	0.1295	0.07	0.02	98.26%
SVD	0.1694	0.1294	0.28	0.04	98.26%

KNNBaseline achieves the best CF performance on coverable rows (RMSE = 0.1666, MAE = 0.1240), outperforming BaselineOnly (−0.0015 RMSE) and SVD (−0.0028 RMSE). All CF models share identical coverage since coverability depends on user/item presence in training, not model type. KNNBaseline is selected for ensemble integration.

5.5 Overall Ensemble Comparison

Ensemble strategies are employed to integrate predictions from feature-based and CF models in order to address varying levels of data availability. Feature-based models provide full prediction coverage across all student–course pairs, while CF models offer stronger personalization under warm-start conditions where sufficient interaction data exist. The objective of this comparison is to evaluate whether ensemble integration can achieve a practical balance between these two properties, improving accuracy over feature-only models while maintaining applicability across the full dataset. Two integration strategies are evaluated: Hybrid-A, which applies deterministic routing to preserve CF predictions when available and fall back to feature-

based predictions otherwise, and Hybrid-B, which combines both signals through weighted blending with a validation-tuned parameter. Both strategies are compared against their component models on the same evaluation subsets to ensure a fair and interpretable assessment of the contribution of ensemble integration.

Table 5. Overall model comparison (VAL and TEST).

Model	Split	N	RMSE	MAE	Coverage
Hybrid-B ($\alpha = 0.65$)	VAL	13,698	0.1619	0.1224	Full (100%)
CF KNN (coverable only)	VAL	13,460	0.1666	0.1240	Partial (98.3%)
Hybrid-A (Routing)	VAL	13,698	0.1668	0.1244	Full (100%)
Feature FE+Sim	VAL	13,698	0.1761	0.1360	Full (100%)
Hybrid-B ($\alpha = 0.65$)	TEST	13,205	0.1684	0.1254	Full (100%)
CF KNN (coverable only)	TEST	12,131	see note†	see note†	Partial (91.9%)
Hybrid-A (Routing)	TEST	13,205	0.1730	0.1261	Full (100%)
Feature FE+Sim	TEST	13,205	0.1777	0.1364	Full (100%)

Hybrid-B achieves the best full-coverage regression on both VAL (RMSE = 0.1619) and TEST (RMSE = 0.1684). The blending parameter α was selected via a sweep over $\alpha \in \{0.00, 0.60, 0.65, 0.70, 1.00\}$ on validation data only; the results are presented in Table 6. The optimal $\alpha = 0.65$ reflects the dataset structure: CF provides stronger personalization when interaction data are available, while feature-based predictions supply stability and coverage. This value was selected because student-level features in the dataset are limited to historical aggregates and do not fully capture individual learning patterns, so CF contributes the stronger personalization signal while feature predictions stabilize results in sparse cases. Performance rankings are consistent across both temporal splits, providing indirect evidence of generalization.

Hybrid-A (Routing) achieves RMSE = 0.1668 on VAL and 0.1730 on TEST, closely matching CF on coverable rows while extending coverage to 100% of instances. The performance difference between Hybrid-A and Hybrid-B on VAL (0.1668 vs. 0.1619, −0.0049 RMSE) highlights the value of partial signal integration: even when CF interaction data are available, blending with feature-based predictions tends to improve accuracy, particularly for near-cold users whose CF neighborhood relationships may be weak. Hybrid-A's advantage lies in interpretability, since every prediction is derived from a single model under a well-defined condition, making it a suitable choice in deployment scenarios where explainability is prioritized over marginal accuracy gains.

The sweep results confirm that intermediate blending ($\alpha = 0.65$) consistently outperforms both extremes: CF-only blending ($\alpha = 1.00$, RMSE = 0.166837 on coverable rows, equivalent to Hybrid-A routing) and Feature-only ($\alpha = 0.00$, RMSE = 0.176122). The gain from $\alpha = 0.65$ over $\alpha = 0.60$ and $\alpha = 0.70$ is small in absolute terms (RMSE differences of 0.000029 and 0.000156 respectively), indicating

that the system is not highly sensitive to precise α selection within the [0.60, 0.70] range. However, the monotonic ordering within this range, confirmed on both VAL RMSE and MAE, justifies the selection of 0.65 as the most conservative choice. The parameter is frozen from validation before test evaluation, ensuring that no information from the test set influences the final configuration.

Table 6. Hybrid-B blending parameter (α) tuning results (VAL).

α	RMSE (VAL)	MAE (VAL)
0.00 (Feature-only)	0.176122	0.135971
0.60	0.161965	0.122720
0.65 (selected)	0.161936	0.122436
0.70	0.162092	0.122309
1.00 (CF-only on coverable)	0.166837	0.124411

5.6 User-Group Performance

To examine model behavior under varying data availability conditions, performance is reported separately for two user groups defined from training data only. Warm users are students for whom CF can generate predictions, meaning both the student and at least one relevant course appear in the training data with sufficient interaction history. Near-cold users are CF-coverable students whose support-set interaction count falls at or below the first quartile threshold (K = 10), representing students who have some enrollment history but not enough for CF neighborhood relationships to be fully reliable. This distinction is important because aggregate metrics alone can obscure systematic performance differences: improvements observed on dense-data subsets may mask degradation for students with limited records, which is precisely the scenario where a deployable recommender system must remain robust.

The near-cold group is particularly important from a practical standpoint. In academic environments, many students are in early stages of their program and have completed only a small number of courses at the time of prediction. For these students, CF cannot reliably identify similar peers because the overlap between their enrollment history and those of other students in the training data is limited. Feature-based models can still generate predictions using course-level statistics and semantic similarity, but they cannot exploit the fine-grained behavioral signals that CF captures for students with richer histories. This is precisely the regime where ensemble blending is expected to add the most value: by combining the available but noisy CF signal with the more stable feature-based prediction, the blended output can partially compensate for the weakness of each component model.

It is also worth noting that the near-cold threshold K = 10 was derived from the empirical distribution of support-set interaction counts and corresponds to the first quartile of that distribution. A sensitivity analysis over $K \in \{2, 3, 5, 10\}$ confirmed that the relative ordering of model performance across user groups remained consistent

across all threshold values, indicating that the reported near-cold results are not an artifact of a particular threshold choice. The warm group, by contrast, consists of students with richer interaction histories for whom CF neighborhood relationships are expected to be more stable and informative. Reporting performance separately for these two groups, using identical data splits and subgroup masks applied consistently across all compared models, ensures that the evaluation reflects genuine differences in modeling effectiveness rather than differences in evaluation methodology. Table 7 reports regression performance across both user groups on VAL and TEST, enabling a direct assessment of whether ensemble gains for near-cold users come at the expense of warm-user accuracy.

Table 7. Regression performance by user group (VAL and TEST).

Model	Split	Group	N	RMSE	MAE	Coverage
Hybrid-B	VAL	NearCold	5,043	0.1653	0.1253	Full
Hybrid-A	VAL	NearCold	5,043	0.1711	0.1306	Full
CF KNN	VAL	NearCold	5,016	0.1870	0.1416	Partial
Feature FE+Sim	VAL	NearCold	5,043	0.1916	0.1497	Full
Hybrid-B	VAL	Warm	8,655	0.1528	0.1154	Full
CF KNN	VAL	Warm	8,444	0.1533	0.1136	Partial
Feature FE+Sim	VAL	Warm	8,655	0.1664	0.1280	Full
Hybrid-A	VAL	Warm	8,655	0.1539	0.1144	Full
Hybrid-B	TEST	NearCold	5,039	0.1938	0.1503	Full
Hybrid-A	TEST	NearCold	5,039	0.2047	0.1513	Full
Feature FE+Sim	TEST	NearCold	5,039	0.2230	0.1772	Full
Hybrid-B	TEST	Warm	8,153	0.1684	0.1254	Full
CF KNN	TEST	Warm	7,649	see note†	see note†	Partial
Hybrid-A	TEST	Warm	8,153	0.1730	0.1261	Full
Feature FE+Sim	TEST	Warm	8,153	0.1777	0.1364	Full

On VAL, Hybrid-B reduces NearCold RMSE from 0.1916 (Feature) to 0.1653 (−13.7%); Hybrid-A achieves 0.1711 (−8.6%). On TEST, Hybrid-B achieves 0.1938 vs. 0.2230 for Feature (−13.1%) and 0.2047 for Hybrid-A (−5.3%). Converting to GPA scale (×4.0): Hybrid-B achieves ±0.78 vs. ±0.89 for Feature-only on TEST NearCold. For Warm users, Hybrid-B (VAL 0.1528, TEST 0.1684) is closely competitive with CF (VAL 0.1533) while providing full coverage, suggesting that blending does not substantially trade warm-start accuracy for near-cold gains. These results indicate why Hybrid-B is particularly suited to near-cold users: CF provides useful but potentially noisy interaction signals when neighborhood relationships are weak, and combining these with feature-based estimates through blending stabilizes predictions in a way that routing alone cannot achieve. The selected blending weight $\alpha = 0.65$ reflects this

balance directly: CF contributes the stronger personalization signal when partial interaction data are available, while the feature-based component supplies stability and ensures coverage when CF signals are unreliable or absent. Improvement is consistent across both temporal splits and six metrics (RMSE, MAE, ACC, BAL_ACC, F1-macro, F1-weighted), providing indirect evidence of robustness, though formal significance testing has not been conducted.

Table 8. Classification performance (Recommend / Neutral / Not Recommend) on VAL and TEST.

Model	Split	N	ACC	BAL_ACC	F1-macro	F1-wt.
Hybrid-B (α=0.65)	VAL	13,698	0.726	0.634	0.638	0.733
Hybrid-A (Routing)	VAL	13,698	0.719	0.630	0.631	0.726
CF KNN (cov. only)	VAL	13,460	0.716	0.629	0.630	0.723
Feature FE+Sim	VAL	13,698	0.714	0.621	0.622	0.724
Hybrid-B (α=0.65)	TEST	13,205	0.689	0.576	0.568	0.703
CF KNN (cov. only)	TEST	12,131	0.700	0.590	0.574	0.709
Hybrid-A (Routing)	TEST	13,205	0.686	0.572	0.562	0.697
Feature FE+Sim	TEST	13,205	0.646	0.534	0.517	0.667

Thresholds t_neutral = 0.52 and t_recommend = 0.73 are tuned on validation and frozen for test. On VAL, Hybrid-B achieves best performance on all four metrics (ACC = 0.726, BAL_ACC = 0.634, F1-macro = 0.638, F1-wt. = 0.733). On TEST, CF achieves ACC = 0.700 on its 91.9% coverable subset; Hybrid-B is the strongest full-coverage model (ACC = 0.689, F1-macro = 0.568). Feature-only shows the largest degradation (F1-macro 0.622→ 0.517), confirming interaction-based signals are essential under temporal distribution shift.

5.8 Item-Group Performance

To complement the user-group analysis, performance is also examined across item groups to assess how well the system handles courses that were not present in the training data. This distinction is particularly important for evaluating the practical scope of the proposed ensemble technique, since newly introduced courses represent a realistic and unavoidable deployment scenario in evolving academic curricula. Seen items are courses that appear in both the training interactions and the test set, for which CF can potentially generate predictions depending on student coverage. New items are courses that appear in the test set but have no prior interaction history in the training data, meaning that CF cannot generate predictions for them by design and all full-coverage models must rely entirely on feature-based representations. Table 9 reports

classification accuracy and, where applicable, regression RMSE for each item group on the test set, providing a direct assessment of the system's generalization capability under new-item conditions.

Table 9. Performance by item group on TEST. †CF regression not available for partial-coverage rows.

Model	Split	Group	N	RMSE	ACC	Coverage
Feature / Hybrid-A / Hybrid-B (tie)	TEST	NewItem	834	—	0.478	Full
Hybrid-B ($\alpha = 0.65$)	TEST	SeenItem	12,371	0.168	0.703	Full
CF KNN (cov. only)	TEST	SeenItem	12,131	see note†	0.700	Partial (91.9%)
Feature FE+Sim	TEST	SeenItem	12,371	0.178	0.658	Full

All full-coverage models are identical for new-item courses (ACC = 0.478) because ensembles fall back to the feature model when CF cannot predict. The sharp drop from VAL (ACC ≈ 0.899) to TEST reflects distribution shift: new courses introduced in later semesters differ in content, difficulty, and grading from the training corpus in ways TF-IDF similarity cannot capture. For seen items, Hybrid-B (ACC = 0.703) outperforms Feature (0.658) and closely matches CF (0.700, partial coverage) with full dataset coverage.

6 Discussion

Three main conclusions follow from the experimental results. First, feature-based models with FE + Sim provide the necessary full-coverage capability. Structured academic aggregates, particularly course difficulty proxies, user historical GPA, and major–course interaction features, are the dominant predictive signals. All learned models substantially outperform the global-mean trivial predictor (RMSE ≈ 0.21–0.24), confirming meaningful information in the available records.

Second, ensemble modeling addresses the CF cold-start limitation while preserving warm-start strengths. Hybrid-B's near-cold gains are consistent across both temporal splits and all six reported metrics, providing consistent empirical support that blending interaction-based and content-based signals yields measurable improvement. The selected $\alpha = 0.65$ reflects that student-level features in this dataset are limited to historical aggregates and do not fully represent individual academic behavior; CF supplies the stronger personalization signal when available, while feature predictions provide robustness and coverage when it is not. Hybrid-B simultaneously maintains near-CF performance for warm users, without substantial warm-user degradation.

Third, new-item generalization remains the primary limitation. Performance drops from ACC ≈ 0.90 on VAL to 0.48 on TEST for new courses, reflecting distribution shift and the inability of TF-IDF similarity to capture course difficulty,

instructor behavior, and assessment design. This defines the system's deployment boundary and motivates future work.

Several threats to validity are acknowledged. The dataset originates from a single institution. Threshold selection introduces policy sensitivity. Formal significance testing (Wilcoxon signed-rank tests, bootstrap confidence intervals) is not reported; however, cross-split and cross-metric consistency across two temporally separated sets and six independent summary statistics substantially reduces the likelihood of spurious findings [18]. The absence of instructor-level and workload-related features is a further limitation. Prediction errors of ±0.78 GPA (Hybrid-B, NearCold, TEST) confirm that the system is best framed as a supplementary decision-support tool, and the three-class categorization is specifically designed to accommodate this uncertainty.

The degree of confidence in the main findings can be assessed through three supplementary lines of evidence, following the framework of Section 5.12 of the original thesis. First, cross-split consistency: the finding that Hybrid-B achieves the best near-cold regression performance among full-coverage models holds on both the validation set (RMSE = 0.1653 vs. 0.1916 for Feature, −0.0263) and the independently held-out test set (RMSE = 0.1938 vs. 0.2230, −0.0292). Observing the same ordering on two temporally separated sets that differ in grade distribution, course composition, and new-item proportion reduces the probability of a spurious result, though it does not eliminate this possibility in the absence of formal significance testing. Second, cross-metric consistency: Hybrid-B's advantage over the feature-only baseline for near-cold users is observed simultaneously across all six reported metrics on both splits, providing indirect evidence against a spurious finding. Third, effect size relative to the trivial baseline: Hybrid-B achieves RMSE = 0.162 overall on validation, a reduction of approximately 24–33% relative to the global-mean predictor (RMSE ≈ 0.21–0.24), confirming that the learned representations add substantial predictive value. To fully establish statistical significance, future work should apply the Wilcoxon signed-rank test on per-record squared errors and construct bootstrap 95% confidence intervals for all RMSE and MAE estimates.

Practical implications for deployment: (1) A feature-based component is essential since new courses and sparse interaction patterns are unavoidable in real curricula. (2) When sufficient interaction data exist, CF provides stronger personalization and should be retained. (3) Ensemble blending is particularly valuable for near-cold students who have some history but not enough for CF alone to be reliable. (4) Decision-level evaluation provides a more interpretable and actionable output than point predictions, and is well-suited for supplementary course-consideration tools.

7 Conclusion

This paper proposes and evaluates an ensemble collaborative filtering technique for an elective courses recommender system at CMU. The technique integrates Ridge regression with FE + Sim features and KNNBaseline CF through deterministic routing (Hybrid-A) and weighted blending (Hybrid-B), evaluated under a time-aware, leakage-

controlled protocol. All parameters and thresholds are frozen from validation before test evaluation.

Three key findings emerge. Hybrid-B achieves the best full-coverage performance on both VAL (RMSE = 0.162, ACC = 0.726) and TEST (RMSE = 0.168, ACC = 0.689). The most pronounced regression improvement is observed for near-cold users: Hybrid-B reduces TEST RMSE by 13% relative to Feature-only (0.1938 vs. 0.2230, or ±0.78 vs. ±0.89 GPA). These gains are consistent across two temporal splits and six metrics. The primary contributions, namely a leakage-controlled evaluation framework with consistent subgroup definitions and systematic near-cold quantification, address weaknesses in prior academic recommendation research.

Future work should prioritize: transformer-based course embeddings (e.g., BERT) to improve new-item generalization; adaptive blending weights varying by user group or workload; instructor-level and section-level contextual features; formal significance testing and multi-institution validation; and separate evaluation by elective category (Major Elective, Free Elective, General Education). The system output should continue to be interpreted as a supplementary tool for selectable course consideration, not a definitive recommendation.

References

[1] Romero, C. and Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. IEEE Trans. Systems, Man, and Cybernetics, Part C, 40(6), 601–618.

[2] Baker, R.S. and Inventado, P.S. (2014). Educational Data Mining and Learning Analytics. In Learning Analytics. Springer, 61–75.

[3] Koren, Y., Bell, R. and Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. Computer, 42(8), 30–37.

[4] Schafer, J.B., Frankowski, D., Herlocker, J. and Sen, S. (2007). Collaborative Filtering Recommender Systems. In The Adaptive Web. Springer, 291–324.

[5] Pazzani, M. and Billsus, D. (2007). Content-Based Recommendation Systems. In The Adaptive Web. Springer, 325–341.

[6] Adomavicius, G. and Tuzhilin, A. (2005). Toward the Next Generation of Recommender Systems. IEEE Trans. Knowledge and Data Engineering, 17(6), 734–749.

[7] Burke, R. (2002). Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction, 12(4), 331–370.

[8] Ricci, F., Rokach, L. and Shapira, B. (2015). Recommender Systems Handbook. 2nd ed. Springer.

[9] Steck, H. (2018). Calibrated Recommendations. In Proc. 12th ACM RecSys. ACM.

[10] Salton, G. and Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval. Information Processing & Management, 24(5), 513–523.

[11] Hastie, T., Tibshirani, R. and Friedman, J. (2009). The Elements of Statistical Learning. 2nd ed. Springer.

[12] Pedregosa, F. et al. (2011). Scikit-Learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[13] Chawla, N.V. et al. (2002). SMOTE: Synthetic Minority Over-sampling Technique. J. Artificial Intelligence Research, 16, 321–357.

[14] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

[15] Friedman, J.H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5), 1189–1232.

[16] He, X. et al. (2017). Neural Collaborative Filtering. In Proc. 26th International Conference on World Wide Web, 173–182.

[17] Hug, N. (2020). Surprise: A Python Library for Recommender Systems. Journal of Open Source Software, 5(52).

[18] Powers, D.M.W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. J. Machine Learning Technologies, 2(1), 37–63.