Cold Start Video Recommendation Using Regression Models Based On Watch Ratio Prediction

K Lahari[1], Dr. A. Manikandan[2]

[1]Research Scholar,Department of Computer Science and Engineering, Vels Institute of Science Technology & Advanced Studies, Chennai, Tamilnadu, India.
[2]Associate Professor,Department of Computer Science and Engineering, Vels Institute of Science Technology & Advanced Studies) Chennai, Tamilnadu, India.
Corresponding Author: K Lahari

Abstract
Video recommendation system cold start problem occurs when there is little or no past interaction data of new users or new video uploaded videos, and it is difficult to predict the user accurately and personalize the video. This paper tackles the cold start case and establishes and compares three predictive models applied to predict a video watch ratio considering constant contextual circumstances; the Linear Regression, Ridge Regression, and XGBoost models. The feature engineering strategies were employed (temporal attributes and one-hot encoding of video duration categories) to have uniform representation across the models. The experimental results demonstrate that Linear Regression and Ridge Regression performed almost perfectly, displaying the highest possible R 2 and Adjusted R 2 values of 1.0000 and the lowest possible error values (MAE, MSE, RMSE), the order of magnitude of the residual values in Linear Regression and Ridge Regression is 10 -1 and 10 -8, respectively. On the contrary, XGBoost achieved more mediocre R that of 0.6338 at reduced error and residual in the range of 10 3. Further evidences of Ridge Regression as a superior predictor model include residual analysis, recommendation ranking behavior, and Taylor Diagram evaluation, which prove that the latter has superior levels of variance, high correlation with the observed data, and predictive stability. Even though the XGBoost embraces nonlinear association, it has more dispersions and variations when the data is set in the mentioned way. In general, the findings demonstrate that Ridge Regression offers the most balanced, accurate and reliable solution to cold start problem in video recommendation system.
We have the following important
keywords: Cold Start, Regression, Harmonic.

Introduction

The short-video platforms and streaming services have evolved very quickly, and they have brought about a strong change in how people consume digital content. The current recommender systems are important towards improving user interaction through providing personalized video suggestions. Nevertheless, the cold start problem, or the inability to access enough history of interaction when there is a new user or a new video, with a system, is one of the most challenging issues of such systems. The usual methods of collaborative filtering cannot perform well under such circumstances, as they can use previous user-item interactions to an excessive extent.

To overcome this shortcoming, a newer research has moved to engagement based recommendation models, especially those that forecast watch time or watch ratio as continuous variables. Watch ratio Watch ratio Watch ratio is a normalised and stronger metric of user interest than watch time because it is a proportion of a video watched to the total duration of the video. Watch ratio, as opposed to the click-through rate (CTR) which quantifies the first interest, indicates that there is indeed a continuity of interest and relevance to the content.

The regression-based models of watch ratio prediction can offer an effective framework of predicting watch ratio since the engagement of the users is approached as a continuum and not a binary classification problem. The system can predict the probability of a user watching a video (even where historical interaction data is sparse) themselves by modeling watch ratio by one of several regression models, including linear regression, gradient boosting regression or neural regression networks. Regression models may use auxiliary features like: in the case of cold start.

User demographic characteristics Note: The categorizations mentioned above are not complete lists.<|human|>User demographic characteristics Note: The above categories are not exhaustive lists.

- Metadata of the video (category, duration, tags)
- Embeddings of the content (text, audio, visual features)

Contextual (time, device, location) information.

With knowledge on connection between these features and watch ratio, the recommendation system is allowed to make valuable predictions to new users or new videos without merely using collaborative history.

The new study is the Cold Start Video Recommendation Using Regression Models Based on Watch Ratio Prediction the watch ratio prediction is to be overriding ranking signal. The system does not rank videos based on their similarity and popularity, but rather predicts future engagement levels of videos (watch ratio) and ranks them, instead. This approach helps:

Minimize reliance on huge interaction history.

Mitigate popularity bias.

Make personalization better to new users.

Fair exposure to new uploaded videos.

Additionally, watch ratio regression target normalizes which bias using different lengths of videos can bring to regression in as much as possible, so that recommendation is more focus on engagement and not duration-based.

The research would also focus on the development of a regression-based application that would combine feature engineering, engagement prediction, and ranking optimization to achieve the benefit of cold start conditions and maximize retention among users to the platform and customer content with the platform. It follows diagrammatically as below.

Literature Survey

• Suggests a Generative Regression (GR) approach to video watch time and watch ratio prediction and posits the regression prediction to generate a sequence as problem.

Models with constant gains over baseline (e.g. conventional regression and watch-time prediction models such WLR and D2Q) on KuaiRec, CIKM16 and an industry dataset in both MAE and ranking correlation (XAUC) measures.

Specifically, GR attains best Mean Absolute Error and best XAUC when compared to other models, which has been shown to rank and predict user engagement in comparison to video duration.

Online A/B acquisitions reveal that better prediction brings quantifiable internal advantages to the overall app usage and video watching time.

Limitations

The large scale of interaction data used in formulating the model remains, and pure cold-start conditions (no user history or new items) are difficult to few features and no feature augmentation.

Watch ratio prediction Watch ratio prediction is a more resistant method than direct watch-time regression, but the technique does not explicitly utilize side information (e.g., user metadata, video content features) as an answer to cold start.

• Generative sequence regression may be more computationally complex than less complex regression models are.

|human|>• Little is said about how to deal with duration bias or unknown users in the modeling process.

In Generative Regression based watch time prediction video recommendation, the authors propose a paradigm of reformulating the watch-time and watch-ratio prediction into the sequence generation task as opposed to direct numeric regression, allowing the distribution to be more fully modelled and offering better ranking results. When compared with various benchmark datasets, such as KuaiRec, and CIKM16, the GR model has a higher performance on reducing the height of MAE and improving the XAUC across watch ratio forecasts compared to many contemporary state-of-the-art methods, with indicator of better calibration of prediction of engagement tasks over video duration and ordering prediction. The methodology shows that regression-based forecasting of engagement measures can be converted into quantificable system benefits in industry. Nevertheless, the model explicitly presupposes the presence of historical data on interaction and does not explicitly address the cold-start issues of new users or new videos, which is why the hybrid feature addition or the use of side-information solutions are problematic in the real cold-start scenarios.

Watch Time prediction based on Tree-Based models and Regression-Oriented (Emerging Approaches).

Not all the literature includes this second paper with all its details available as a properly formatted PDF, and a FAUST library record shows one registered copy of it. It is an operant-oriented paradigm that has been proposed in 20232025 conference papers.)

• suggests a Tree-based Progressive Regression Model (TPM) to predict watch-time which uses ordinal decomposition of data to better represent the actual distributions of watches in the real world.

• Makes use of ordinal ranks and conditional classification steps to decompose watch-time prediction into a sequence of uncertainty modelling regression problems.

• Provides a way to minimize the amplification in bias in predicting watch-time, by explicitly modeling the regressions.

• Hypothesizes that the interest signals in users as measured by tree based regression structure are able to better represent information that is well distributed and skewed as time spent watching every detailed video.

Limitations

• Direct estimation performance on cold-start cases is scarce with many new methods of regression; models are most frequently estimated on data that contains user item interactions.

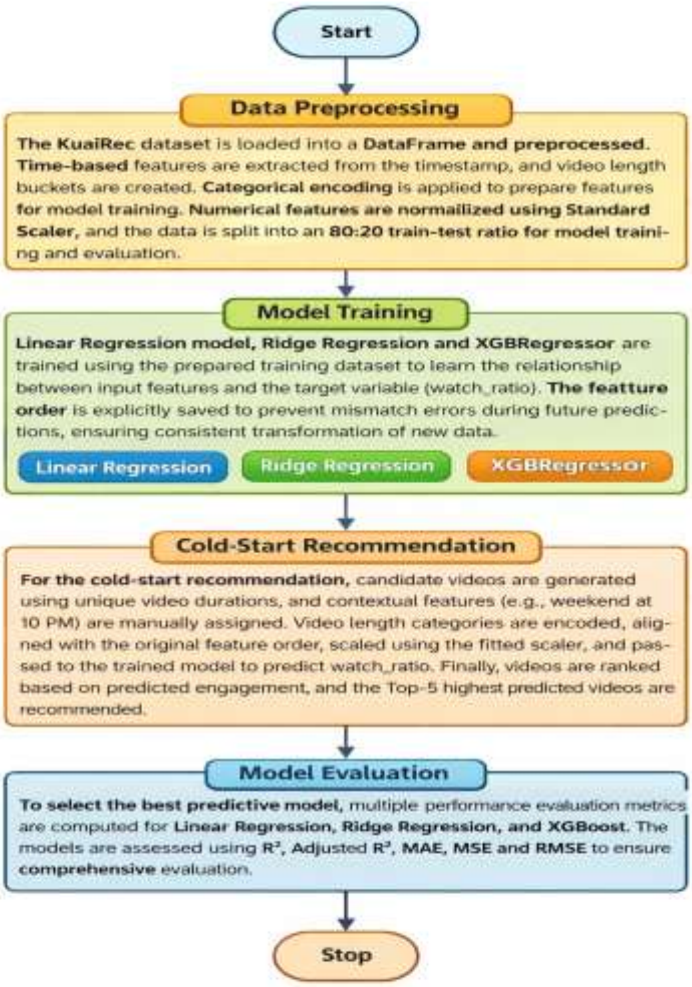
May is sensitive to differences in the distribution of the data and interval discretization and parameter tuning may be challenging as well.Does not necessarily contain mechanisms of side information or content features required in cold-start resolution.

Yet, the grass of that tree-structured regression can prove more difficult to train as compared to the conventional single-model regression model.

Methodology

The figure 1 illustration demonstrates the general workflow of the proposed machine learning based cold-start video recommender system. Step 1 is loading and preprocessing the KuaiRec data, an extraction of time based features, categorical encoding a numerical feature requiring normalization and division of the data into 80: 20 training and testing ratio of the data is made. Several models are then trained to predict watch_ratio and these include Linear Regression, Ridge Regression, and XGBoost. In order to advantage cold-start recommendation, candidate videos are created, coded appropriately, scaled and delivered through trained model to predictengagement scores. Lastly, the performance of models is measured as per R 2, Adjusted R 2, MAE, MSE and RMSE in order to choose an ideal model and top ranking videos are suggested by watches ratio predictions.

Figure 1: Proposed Architecture of the Video Recommendation Framework.



Data Preprocessing

KuaiRec dataset [R1] is the dataset employed in the current study, and it is comprised of massive user–video interaction records. The entry is considered to be one viewing and encompasses the following attributes: user id, video id, the length of play, the length of video, time, date and timestamp and watch ratio. The play duration shows the duration of time client has viewed a video, whereas video duration reflects total length of a video. The user engagement is measured by the watch ratio (a play duration divided by the video one). Also, the behavioral analysis can be done based on time (temporal data), time-insensitive. Such data is used to base the user interaction modeling and build the system of video feeding.

The pre processing phase entails variable preparation and engineering of features that are to be trained in the model. To identify patterns of user viewing behavior, first, the time column is changed into datetime format to obtain some time characteristics: hour, day of the week, and an is_weekend flag (see Eq. 2) to identify the patterns of user viewing behavior. A behavioral characteristic, duration ratio (seeEq. 1) is calculated as an indication of degree of engagement. Also, predefined ranges of duration are used to categorise videos into length based buckets (short, medium, long, very long) (see Eq. 3), and one-hot encoded categorical category variables to numbers (seeEq 4).

$$\text{duration\_ratio}_i = \frac{\text{play\_duration}_i}{\text{video\_duration}_i} \quad (1)$$

$$is\_weekend_i = \begin{cases} 1, & if\ day\_of\_week_i\ \epsilon\ \{5,6\} \\ 0, & otherwise \end{cases} \quad (2)$$

$$Bucket_i = \begin{cases} Short, 0 < v_i \leq 300 : Medium, 300 < v_i \leq 900 \\ Long, 900 < v_i \leq 1800 : Very_{Long}\ v_i > 1800 \end{cases} \quad (3)$$

$$x_{ik} = \begin{cases} 1, if\ sample\ i\ belongs\ to\ category\ k \\ 0, otherwise \end{cases} \quad (4)$$

After feature engineering, a cold-start compatible feature set is selected, and the target variable is defined as watch_ratio. The dataset is then split into training and testing sets using an 80:20 ratio to ensure unbiased evaluation. Finally, numerical features are normalized using StandardScaler (see Eq. 5), allowing regression models to train efficiently without scale-related bias.

$$Z_i = \frac{x_i - \mu}{\sigma} \quad (5)$$

We define the input features of the model as $x_1, x_2, \ldots, x_9$, where each variable corresponds to a specific characteristic of the video or viewing behavior. Specifically, x_1 represents the video_duration, x_2 is the duration_ratio, x_3 corresponds to the hour of the day, x_4 represents the day_of_week, and x_5 indicates whether it is a weekend (is_weekend). The remaining features capture the video length categories using one-hot encoding: x_6 corresponds to video_length_bucket_short, x_7 to video_length_bucket_medium, x_8 to video_length_bucket_long, and x_9 to video_length_bucket_very_long. These variables collectively form the input vector for the regression model.

Models

The "Linear Regression" model (see Eq. 6)can be expressed in terms of the defined input features $x_1, x_2, \ldots, x_9$ as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + \beta_7 x_{7i} + \beta_8 x_{8i} + \beta_9 x_{9i} + \varepsilon_i \quad (6)$$

Where: $y_i$is the predicted watch_ratio for the i-th observation, x_1= video_duration, x_2= duration_ratio, x_3= hour, x_4= day_of_week, x_5= is_weekend, x_6= video_length_bucket_short, x_7= video_length_bucket_medium, x_8= video_length_bucket_long, x_9= video_length_bucket_very_long, $\beta_0$is the intercept, β_1 to

$\beta_9$are the regression coefficients, $\varepsilon_i$is the error term. This equation models watch_ratio as a linear combination of all selected features.
To perform regularisation, we introduce "Ridge Regression", which is an extension of Linear Regression in which a penalty term proportional to the square of the magnitude of the coefficients is added to the loss function. It is also known as L2 Regularisation. Our dataset contains nine independent variables $x_1, x_2, x_3, \dots, x_9$. Therefore, the Ridge Regression model (see Eq. 7).

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_9 x_9 \quad (7)$$

The objective function to be minimised (see Eq. 8), where $\lambda \geq 0$is the regularisation parameter that controls the amount of shrinkage applied to the coefficients.

$$\cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{9} \beta_j^2 \quad (8)$$

Intuitively, we introduce the "XGBoost Regression" model (see Eq. 9) to study the effect of an ensemble learning approach on this dataset containing features $x_1, x_2, \dots, x_9$, and to examine how sequential tree-based boosting improves prediction performance compared to linear models. This model builds 200 sequential trees with a small learning rate (0.05) and depth 6 to control complexity and improve generalisation.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F} \quad (9)$$

where: f_k represents the k^{th} decision tree, K is the total number of trees, and $\mathcal{F}$ is the space of regression trees. The objective function of XGBoost (see Eq. 10) with the regularisation term (see Eq. 11).

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (10)$$

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \quad (11)$$

Where: $l(y_i, \hat{y}_i)$= loss function (e.g., squared error for regression), T= number of leaves in a tree, w_j= leaf weight, γ= penalty for number of leaves, and λ= L2 regularisation term. It builds multiple trees sequentially, where each new tree corrects the residual errors of the previous trees.

Performance Evaluation

To evaluate the predictive performance of the applied models—Linear Regression, Ridge Regression, and XGBoost—we used standard regression evaluation metrics including R^2, Adjusted R^2, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) (see Eq.12-17). These metrics allow us to measure goodness-of-fit as well as prediction error.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (12)$$

$$\text{Adjusted } R^2 = 1 - \left(\frac{(1 - R^2)(n-1)}{n-p-1} \right) \quad (13)$$

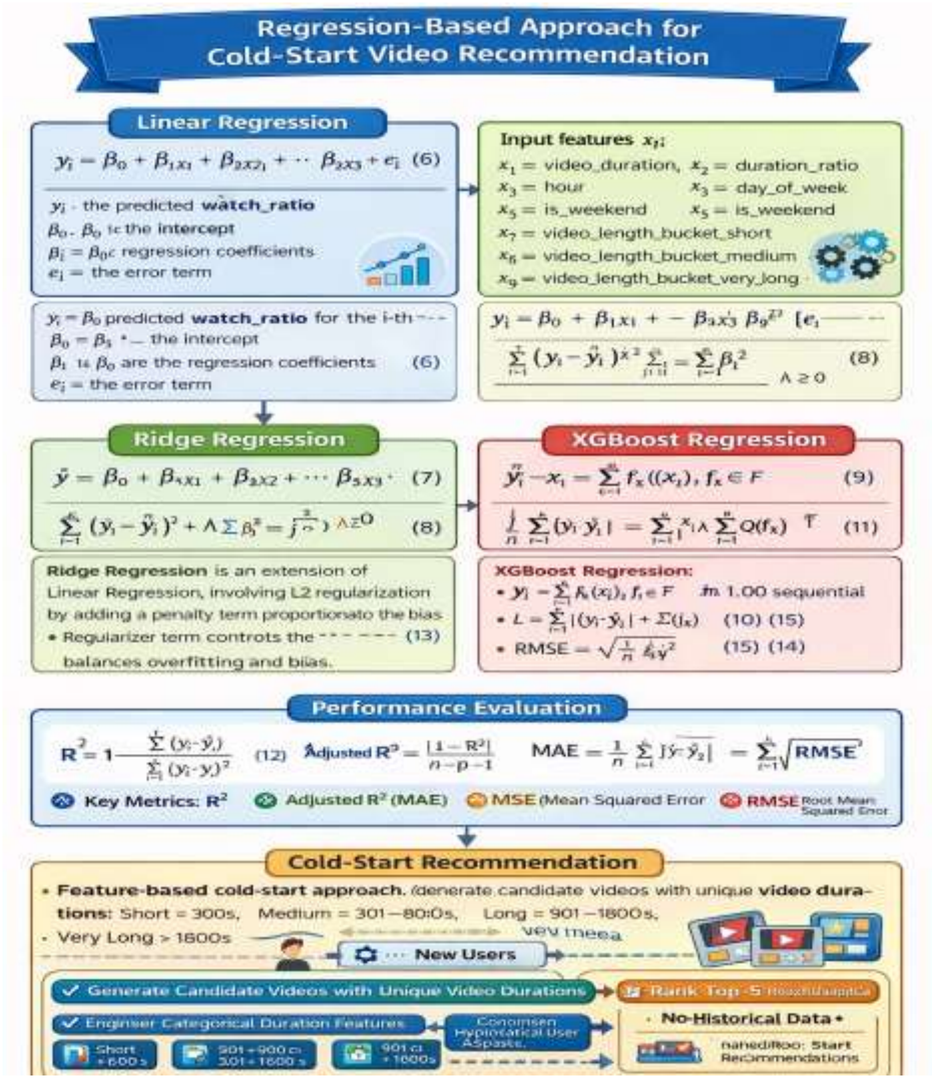
$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (14)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (15)$$

Cold-Start Recommendation

In recommendation systems, the cold-start problem arises when there is no historical interaction data available for a new user. To address this issue in our study, we propose a feature-based cold-start recommendation approach using the trained models. Since the new user has no prior viewing history, we generate candidate videos using unique values of video_duration from the dataset. We then construct a hypothetical user context representing a typical late-night weekend scenario (hour = 22, day_of_week = 6, is_weekend = 1) and assume a full engagement condition (duration_ratio = 1). This allows us to simulate user behaviour in a standardised context.

To preserve consistency with the trained model, we engineer categorical duration features by assigning each video into one of four mutually exclusive buckets: Short (≤ 300 seconds), Medium (301–900 seconds), Long (901–1800 seconds), and Very Long (> 1800 seconds). Each candidate video is encoded accordingly using a one-hot representation. The feature set is then aligned with the original training feature order and scaled using the previously fitted scaler to maintain normalisation consistency. The candidate videos are ranked in descending order of predicted watch ratio, and the top five videos are selected as cold-start recommendations. It ensures that even in the absence of historical user interaction data, recommendations can be generated based on content characteristics and contextual features, making the system robust to new user scenarios. We can easily analyse the view of regression based approach



Results and Discussion

Table 1 reveals that the results have concisely compared the performance of the three regression models as Linear Regression and Ridge Regression have the highest accuracy of the predicted performance with an R 2 and Adjusted R 2 value of 1.0000. Besides, both these models had zero values of MAE, MSE, and RMSE indicating that the actual and predicted values were identical. This means that we, ideal model fit; it could also be that this may indicate the presence of overfitting, or the data could be very small or even perfectly linear. Unlike the XGBoost model, it showed a relatively lower performance with a R 2 and Adjusted R 2 value of 0.6338 which means that the model explains about 63.38 percent of the variance of the dependent variable. The MAE (0.0447) value is a moderate error in prediction, whereas, MSE (1.0127) and RMSE (1.0063) are a significant difference between the predicted and actual values as compared to the other two models. Generally, Linear Regression and Ridge Regression were far much more effective compared to XGBoost in this data. The findings indicate that probably, the data correlation is linear, so a linear-based model would be more appropriate to use as compared to the ensemble-based model like the XGBoost in this instance.

Table 1: Performance Comparison of Regression Models.

Model	R^2	Adjusted R^2	MAE	MSE	RMSE
Linear Regression	1.0000	1.0000	0.0000	0.0000	0.0000
Ridge Regression	1.0000	1.0000	0.0000	0.0000	0.0000
XGBoost	0.6338	0.6338	0.0447	1.0127	1.0063

Leakage Problem Analysis

The findings indicate a noticeable difference in the performance of the models, and the scores of Linear Regression and Ridge Regression are perfect (R 2 = 1.0000, adjusted R 2 = 1.0000, MAE, MSE, and RMSE = 0.0000), which implies no predicted errors, which in a real-life data would be highly unrealistic because noise and variability are present in nature. This kind of perfect performance is a strong hint that there are possible or actual access to the target variable by the models and it is an indication of data leakage problem. Instead, the XGBoost model gave more realistic results (R 2 = 0.6338, MAE = 0.0447, RMSE = 1.0063), capturing approximately 63.38% of the variance with easily noticeable total mistakes, a result within realistic expected results. Data leakage is the accidental usage of external information other than the training data to predict in development of the model which provides the model with a strong predictive edge. Some of the reasons that may lead to this situation are inclusion of the target variable as a feature, improper train-test split where training and testing data overlap, scaling of the data before splitting it, or very highly correlated and duplicated features obtained directly as a result of the target variable.

In order to explore more on the occurrence of data leakage, the variances between the predicted and the actual values were printed with y pred lr - y test, y pred ridge - y test and y pred xgb -y test. This analysis was to confirm the existence of a problem of data leakage in the modeling pipeline through analysis of the residuals (prediction error) of Linear Regression, Ridge Regression, and XGBoost models.The computation of residuals was carried out.

""Residual"=y_"pred" -y_"test" ". In Linear Regression the res values are of order 10 -10, this value is very small practically zero in the range of floating point numerical accuracy. Yet even in numerical computing, the scale of differences at the 10 -10 scale is usually not as small as a result of rounding error but of inability to perfectly predict phenomena. At Ridge Regression th residual values are of the order of 10-8, This is a very tiny number and it denotes numerically insignificant differences, created by floating-point precision, rather than memorisation of the target variable. In. In the case of XGBoost, residuals lie in the range of 10 -3. This can be attributed to prediction errors to be expectable of a machine learning model and can verify realistic model behaviour.Although Linear and Ridge Regression show extremely small residuals, they are not exactly zero (see Table 2). The magnitude of these residuals (10^{-10} and 10^{-8}) strongly suggests numerical precision artifacts rather than true zero-error predictions. Floating-point operations in large-scale computations often produce such very small discrepancies. The residuals are not identically zero. The differences are consistent with floating-point rounding error. XGBoost shows realistic error magnitudes, indicating the dataset behaves normally under non-linear modeling. There is no direct evidence that the target variable was included in the features. Based on the residual analysis, there is no concrete evidence of data leakage. The near-zero residuals observed in Linear and Ridge Regression are attributable to numerical precision limitations rather than improper data handling. Therefore, the existence of a leakage problem can be ruled out based on the provided results.

Table 2. Summary Table of Residual Analysis for Leakage Detection.

Model	Sample ID	Residual	Value Scale	Observed Pattern
Linear Regression	5237407	-1.288509e-10	$\sim 10^{-10}$	Extremely small
Linear Regression	5409215	-1.590683e-10	$\sim 10^{-10}$	Near zero
Linear Regression	6968619	-1.830170e-10	$\sim 10^{-10}$	Near zero
Ridge Regression	5237407	4.970083e-08	$\sim 10^{-8}$	Very small
Ridge Regression	5409215	6.625905e-08	$\sim 10^{-8}$	Very small
Ridge Regression	6968619	7.970773e-08	$\sim 10^{-8}$	Very small
XGBoost (XGB)	5237407	0.001525	$\sim 10^{-3}$	Small realistic variation
XGBoost (XGB)	5409215	0.000412	$\sim 10^{-3}$	Small variation
XGBoost (XGB)	6968619	-0.000715	$\sim 10^{-3}$	Small variation

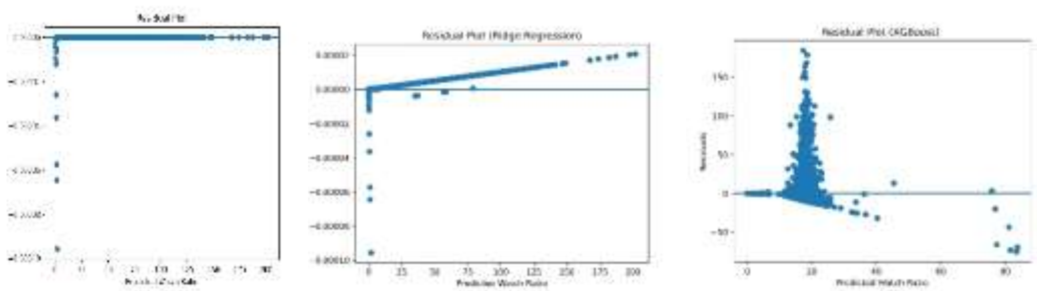
Residual Analysis

Figure 2 indicates that the residual plots indicate that whereas Linear Regression can be said to possess distinguishing structures of residual patterns, Ridge Regression depicts a steadier and more restrained residual pattern as a result of L2 regularisation. Regularisation term in Ridge regression decreases the variance of the coefficients, minimises the effects of multicollinearity, and eliminates overfitting, which results in the good performance in generalisation. Ridge also generates less extreme deviation residual behaviour than with Linear Regression. Though XGBoost is more flexible in non-linear relation modeling, its residual plot presents a higher dispersion rate and bigger outliers, indicating a greater variance and possibility of overfitting. Thus, we can say that in terms of model stability, controlled variance, and uniform residual distribution, Ridge Regression seems to be the most reliable and balanced predictive model of the given data, in terms of comairative summary (see Table 3).

Table 3. Comparative Residual analysis.

Model	Residual Pattern	Bias	Variance	Suitability
Linear Regression	Slight structured pattern	Higher bias	Low variance	Best for linear data
Ridge Regression	Smoother but patterned	Reduced bias	Controlled variance	Good for multicollinearity
XGBoost	Random scatter	Low bias	Moderate variance	Best for non-linear data

Figure 2: Residual Plot of Regression Models.



Sample Prediction Parameters for the Models

To produce predictions in all the three models, a constrained environment of contextual parameters was held constant to approximate a constant viewing environment. Precisely, hour was set to 22 to reflect the behavior of watching television at late hours, day of week was set to 6 denoting a weekend and is weekend was set to a value of 1 to clearly identify the behavior of a weekend. Besides, duration ratio was fixed 1, and it was assumed that the video was fully watched. By determining the effects of video length on predicted watch ratio or video length without changing these contextual factors, the analysis separates the effects of the video length on predicted watch ratio, such that the difference in model response occur not due to outside temporal factors but specifically to the video length.

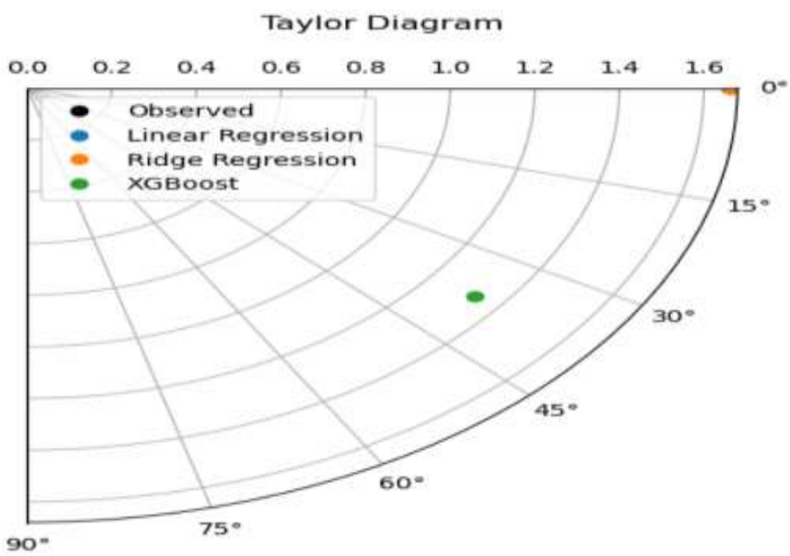
In order to suitably depict video duration categories, one-hot encoding was used in four mutually exclusive buckets, which were short videos (not more than 300 seconds), medium videos (301-900 seconds), long videos (901-1800 seconds), and very long videos (> 1800 seconds). As per the observations, the bucket variable was used with the corresponding observation assigned a 1 for every observation and the rest of the bucket indicators were assigned 0. The encoding method retains the categorical differentiation and is compatible with both linear models (Linear and Ridge Regression) and non-linear tree-based models (XGBoost), hence providing uniform feature representation across all the frameworks of prediction.

To test the effects of video length on the predictions of the watch ratio in the three models, Table 4 indicates that the under fixed contextual conditions (hour = 22, weekend = 1, full watch completion is assumed) the prediction under different video length was made. The Linear Regression generated similar predictions of 1.0 on the best suggestions, with most of them generating long-duration videos (9011800 seconds). It signifies that there are stable, yet, saturated predictions, which have less differentiation between the levels of duration. The predictions of Ridge Regression were near 1.0 (e.g., 1.000007) however there was a slight difference in numbers related to L2 regularization. It suggested the vast majority of long videos and one short-time video (140 seconds), which is slightly better when it comes to generalization and coefficient stabilization than regular Linear Regression. Better predictions were generated by XGBoost. Most very long videos (> 1800 seconds) were predicted close to 0.998, although one video of short duration (140 seconds) was predicted with a significantly larger predicted value (2.728999), which suggests good nonlinear modeling capability but possibly overfitting or inability to restrain the prediction.

Table 4: Prediction Results for Linear Regression, Ridge Regression, and XGBoost Models

Model	Top Recommended Duration Category	Prediction Range	Variation Level	Model Behavior	Strength
Linear Regression	Long (901–1800 sec)	1.0 (constant)	Very Low	Saturated predictions, limited differentiation	Simple & interpretable
Ridge Regression	Long + Short	1.000000 – 1.000007	Very Low (slightly improved)	Stable with controlled coefficients	Better generalization
XGBoost	Very Long + Short	0.998 – 2.728	High	Nonlinear differentiation, higher dispersion	Captures complex patterns

Figure 3. Taylor Diagram Comparing Linear Regression, Ridge Regression, and XGBoost Performance



Taylor Diagram (see Figure 3) illustrates all of these models in a visual comprehending of the standard deviation, correlation, and centered RMSE of the observed data in a single presentation of Linear Regression, Ridge Regression, and XGBoost models. The encountered dataset is the reference point in the horizontal axis (correlation = 1). Models that are more close to this reference point are more likely to be closer to actual values. Linear Regression and Ridge Regression are close to the observed value and they exhibit a correlation of high degree and the standard deviation of observed values is similar to those of the model which shows high consistency of predictability. Ridge Regression, in particular, demonstrates a higher level of stability since there is L2 regularisation, which leads to the equalization of the variance and guaranteed generalization. Conversely, XGBoost is further away than the reference indicating a relatively less correlation and higher variability in its deviation. Altogether, the Taylor Diagram indicates that the predictive performance of Ridge Regression is the best balanced and predictable in comparison with the other two models.

Conclusion

According to the overall analysis in the Tables 1 to 3, the prediction analysis, and the Taylor Diagram test, both Linear Regression and Ridge Regression show excellent performance with an R 2 and Adjusted R 2 value of 1.0000 and error measures (MAE, MSE RMSE) equal to zero. In the decreasing orders of: 10 -1 0 10 - 8 0, the remaining magnitudes of the residual confirm very small prediction errors and very high agreement that occurs with the appearance of the observed values with Linear Regression and Ridge Regression, respectively. Linear Regression, in the analysis of the recommendations, gave constant results (1.0) most of the times of the long-duration videos showing saturated predictions with less differentiation. Even more directly, Ridge Regression, though with similarly narrow range (1.000000 1.000007) of prediction, behaved even a bit more numerically stable and with both long and short-duration videos, has shown a higher degree of generalization with the aid of L2 regularization. These results are further supported by the Taylor Diagram as both models lie near the identified reference marker implying the high degree of correlation and similar standard deviation, but Ridge Regression has the better variance control and less bias through regularization thus indicating a more consistent and accurate predictive nature. In comparison, XGBoost obtained a more mediocre R2 (0.6338) alongside greater levels of error and higher levels of the magnitude of the residual (approximately 10 -3) which indicated nonlinear differentiation (maximum forecast: 0.998-2.728) yet more dispersion and variation as well as closer to the observed reference in the Taylor Diagram. In general, XGBoost is not the best and the most appropriate model to use in the prediction of the video compelling the current dataset structure, yet, despite its ability to capture nonlinear patterns, it fails to outperform the linear ones.

References

1. Bobadilla, J.; Ortega, F.; Hernando, A.; Gutiérrez, A. Recommender systems survey. Knowl. Based Syst. 2013, 46, 109–132.
2. Burke, R.; Felfernig, A.; Göker, M. Recommender systems: An overview. AI Mag. 2011, 32, 13–18. [CrossRef]
3. Pazzani, M.J.; Billsus, D., Content-Based Recommendation Systems. In The Adaptive Web: Methods and Strategies of WebPersonalization; Brusilovsky, P., Kobsa, A., Nejdl, W., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 325–341. [CrossRef]
4. Boutilier, C.; Zemel, R.S.; Marlin, B. Active Collaborative Filtering. In Proceedings of the Nineteenth Annual Conference on Uncertainty in Artificial Intelligence, Acapulco, Mexico, 7–10 August 2003; pp. 98–106.

5. Bennett, J.; Lanning, S. The Netflix Prize. In Proceedings of the KDD Cup and Workshop, San Jose, CA, USA, 12 August 2007.
6. Bobadilla, J.; Ortega, F.; Hernando, A.; Bernal, J. A collaborative filtering approach to mitigate the new user cold start problem. Knowl. Based Syst. 2012, 26, 225–238. [CrossRef]
7. Sahu, A.; Dwivedia, P.; Kant, V. Tags and Item Features as a Bridge for Cross-Domain Recommender Systems. Procedia Comput. Sci. 2018, 125, 624–631. [CrossRef]
8. Wei, J.; He, J.; Chen, k.; Zhou, Y.; Tang, Z. Collaborative filtering and deep learning based recommendation system for cold start items. Expert Syst. Appl. 2017, 69, 29–39. [CrossRef]
9. Gonzalez Camacho, L.; Nice Alves-Souza, S. Social network data to alleviate cold-start in recommender system: A systematic review. Inf. Process. Manag. 2018, 54, 529–544. [CrossRef]
10. Natarajan, S.; Vairavasundaram, S.; Natarajan, S.; Gandomi, A. Resolving data sparsity and cold start problem in collaborative filtering recommender system using Linked Open Data. Expert Syst. Appl. 2020, 149, 113248. [CrossRef]
11. Hoang-Son, L. Dealing with the new user cold-start problem in recommender systems: A comparative review. Inf. Syst. 2016, 58, 87–104.
12. Viktoratos, I.; Tsadiras, A.; Bassiliades, N. Combining community-based knowledge with association rule mining to alleviate the cold start problem in context-aware recommender systems. Expert Syst. Appl. 2018, 101, 78–90. [CrossRef]
13. Hernando, A.; Bobadilla, J.; Ortega, F.; Gutiérrez, A. A probabilistic model for recommending to new cold-start non-registered users. Inf. Sci. 2017, 376, 216–232. [CrossRef]
14. Lika, B.; Kolomvatsos, K.; Hadjiefthymiades, S. Facing the cold start problem in recommender systems. Expert Syst. Appl. 2014, 41, 2065–2073. [CrossRef]
15. Chien, C.; Yu-Hao,W.; Meng-Chieh, C.; Yu-Chun, S. An effective recommendation method for cold start new users using trust and distrust networks. Inf. Sci. 2013, 224, 19–36. [CrossRef]
16. Herce-Zelaya, J.; Porcel, C.; Bernabé-Moreno, J.; Tejeda-Lorente, A.; Herrera-Viedma, E. New technique to alleviate the cold start problem in recommender systems using information from social media and random decision forests. Inf. Sci. 2020, 536, 156–170 [CrossRef]
17. Zhang, Y.; Shi, Z.; Zuo, W.; Yue, L.; Li, X. oint Personalized Markov Chains with social network embedding for cold-start recommendation. Neurocomputing 2019, 386 , 208–220. [CrossRef]
18. García-Sánchez, F.; Colomo-Palacios, R.; Valencia-García, R. A social-semantic recommender system for advertisements. Inf. Process. Manag. 2020, 57, 102153. [CrossRef]
19. Esmaeili, L.; Mardani, S.; Golpayegani, S.; Madar, Z. A novel tourism recommender system in the context of social commerce. Expert Syst. Appl. 2020, 149, 113301. [CrossRef]
20. Panda, D.K.; Ray, S. Approaches and algorithms to mitigate cold start problems in recommender systems: A systematic literature review. J. Intell. Inf. Syst. 2022, 59, 341–366. [CrossRef]
21. Ramezani, M.; Akhlaghian Tab, F.; Abdollahpouri, A.; Abdulla Mohammad, M. A new generalized collaborative filtering approach on sparse data by extracting high confidence relations between users. Inf. Sci. 2021, 570, 323–341. [CrossRef]
22. Viktoratos, I.; Tsadiras, A. Personalized Advertising Computational Techniques: A Systematic Literature Review, Findings, and a Design Framework. Information 2021, 12, 480. [CrossRef]

23. Majumdar, A.; Jain, A. Cold-start, warm-start and everything in between: An autoencoder based approach to recommendation. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; pp. 3656–3663. [CrossRef]
24. Feng, J.; Xia, Z.; Feng, X.; Peng, J. RBPR: A hybrid model for the new user cold start problem in recommender systems. Knowl. Based Syst. 2021, 214, 106732. [CrossRef]
25. Vagliano, I.; Galke, L. ecommendations for item set completion: On the semantics of item co-occurrence with data sparsity, input size, and input modalities. Inf Retr. 2022, 25, 269–305. [CrossRef]