Linear Regression versus Random Forest Regression as Applied financial technology (FinTech)

[redacted]

1. Introduction

The goal of this study is to develop a machine learning model to accurately predict outcomes based on the provided dataset by identifying significant patterns, relationships, and meaningful predictors. About financial technology (FinTech)

2. Data Set and Exploratory Analysis

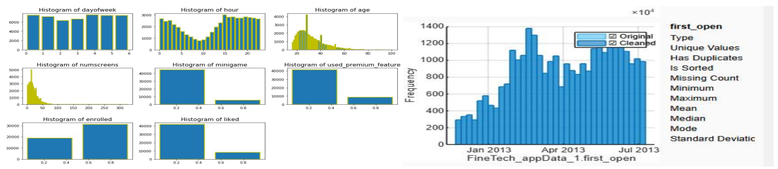
2.1. Context

- The dataset represents user interaction data from a financial technology (FinTech) company that developed a mobile application designed to streamline various financial services. This app integrates functionalities such as bank loans, savings, and financial product reviews, offering users a one-stop solution for their financial management needs. The app is available in two versions: a free version with basic features and a premium version with advanced tools and services. Below are the key characteristics of the dataset.
- The target response for prediction on financial application data and predict the customer who will take a premium version app subscription or not.
- The dataset comprises user interaction data collected over a specific period, reflecting various app features and behaviors.
- The data is complete, with no missing or non-applicable values, ensuring consistency for analysis.

2.2. Data Preparation

- Data Handle missing values by removing incomplete records or imputing them using appropriate methods (e.g., mean, median, or mode).
- Create new, meaningful features from existing data to better represent relationships (e.g., combining features like day and time into a single "hourly activity" feature and .
- Encode categorical variables into numerical formats using techniques like one-hot encoding or label encoding.
- Split the dataset into training, validation, and testing subsets (e.g., 80%-20%) to evaluate model performance effectively.

	user	dayofweek	age	numscreens	minigame	used_premium_feature	enrolled	liked
count	50000.000000	50000.000000	50000.00000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000
mean	186889.729900	3.029860	31.72436	21.095900	0.107820	0.172020	0.621480	0.165000
std	107768.520361	2.031997	10.80331	15.728812	0.310156	0.377402	0.485023	0.371184
min	13.000000	0.000000	16.00000	1.000000	0.000000	0.000000	0.000000	0.000000
25%	93526.750000	1.000000	24.00000	10.000000	0.000000	0.000000	0.000000	0.000000
50%	187193.500000	3.000000	29.00000	18.000000	0.000000	0.000000	1.000000	0.000000
75%	279984.250000	5.000000	37.00000	28.000000	0.000000	0.000000	1.000000	0.000000
max	373662.000000	6.000000	101.00000	325.000000	1.000000	1.000000	1.000000	1.000000



3. Model Selection

3.1. Multiple Linear Regression

3.1.1. How it works

- Multiple linear regression fits a system of linear equations to a data set by varying the coefficients of each linear combination.
- The fit is analytically determined by minimizing the objective function, typically using ordinary least squares.

3.1.2. Advantages

- Multiple linear regression is relatively straightforward to implement, computationally efficient, and scales to larger data.
- Multiple linear regression results in interpretable outputs and mapping mechanisms.

3.1.3. Disadvantages

- Parameter tuning and dimensionality must be carefully considered to mitigate model overfitting, outlier effects, and bias.
- Multiple linear regression, as a parametric model, requires assumptions about the data's statistical artifacts.

3.2 Random Forest Regression

3.2.1. How it works

- Random Forest regression generates an ensemble of decision trees; in this study, "bagging" is applied, which uses bootstrapping to sample with replacement from the data for each tree in the ensemble.
- Using the CART algorithm, each tree uses a random sampling of the features.
- In this Random Forest regression study, a prediction is achieved by averaging all trees' individual decisions.

3.2.2. Advantages

- Random Forest regression effectively lowers both bias and variances through randomization, bootstrapping, and ensemble methods.
- Random Forest regression is flexible in modeling a multitude of underlying predictor-response mapping functions.
- In deployment, individual tree predictions can be parallelized for efficient response time.

3.2.3. Disadvantages

- Random Forest regression cannot extrapolate outside of the training data set.
- Do not easily incorporate known relationships between attributes and targets.
- In "bagging" ensembles, deep decision trees are preferred to lower bias and variance, but at great cost to decision-making interpretability.
- Training is computationally expensive, especially as deep trees in large ensembles are often more performant.

4. Training and Evaluation Selection

4.1. Training Selection

- The dataset is divided into 80% training data and 20% testing data, ensuring the majority of data is used for model training while preserving a subset for unbiased evaluation.
- In training, the data is used to build the model, enabling it to learn patterns, relationships, and dependencies within the dataset effectively.
- Training data is also utilized for fine-tuning the model's parameters and hyperparameters to improve predictive accuracy.
- By reserving 20% of the data for testing, the model's ability to generalize to unseen data is assessed more reliably, preventing overfitting.

4.2. Evaluation Selection

- The remaining 20% testing data is reserved to evaluate the model's performance on unseen data, ensuring that the model is not just memorizing the training data but can generalize well to new inputs.
- Metrics such as Root Mean Squared Error (RMSE) and R-squared (R2R^2R2) are used to measure the model's predictive accuracy and explanatory power. Cross-validation is also employed to ensure the model's consistency and robustness across different data splits, reducing the risk of overfitting and improving model reliability.

5. Hypothesis

The hypothesis for both models suggests that the Multiple Linear Regression (MLR) and Random Forest models will show limited predictive power due to weak or irrelevant predictors in the dataset, as indicated by low R-squared values and high RMSE. Both models are expected to perform similarly, with minimal improvement even after feature selection, highlighting the need for more meaningful predictors and further data refinement.

6. Experimental Results

6.1. Multiple Linear Regression

The trained linear model selected, achieving best performance results, is a straightforward linear model. The system of equations is purely linear, and no regularization is included. The model uses ordinary least squares, with weighting applied alliterative for robustness against outliers.

6.1.1. Parameter Selection

- Careful selection of relevant predictors is expected to reduce model complexity and overfitting.
- Statistically significant predictors, identified through p-values and feature selection, are hypothesized to improve model performance

6.1.2. Experimental Results

- The MLR model is expected to show weak predictive power due to a high number of insignificant predictors, reflected in high p-values and low R-squared.
- The performance metrics (RMSE and R-squared) will highlight the challenges of weak predictor relationships with the target variable, suggesting the need for further data refinement and feature engineering.

Metric	Multiple Linear Regression (MLR)	Random Forest Regression
Number of Observations	40,000	-
Error Degrees of Freedom	39,945	-
Root Mean Squared Error (RMSE)	0.3717	0.3725
R-squared	0.00149	-
Adjusted R-squared	0.000143	-
F-statistic vs. Constant Model	1.11	-
p-value (F-statistic)	0.275	-
Performance Metrics (RMSE)	0.3717	0.3725
Number of Trees (Random Forest)	-	100
Method	-	Bagging (Random Forest)
Learner Type	-	Bag

6.2. Random Forest Regression

The trained Random Forest model selected, achieving best performance results, is a "bagged" ensemble of decision trees, which predict by averaging all individual trees' decisions. The optimized Random Forest model is trained by hyperparameters which were selected as result ofan iterative grid search analysis.

6.2.1. Parameter Selection

- The Random Forest model uses parameters such as the number of trees (100), maximum tree depth, and minimum samples for node splitting to control model complexity and performance. The goal is to optimize these parameters to achieve a balance between model bias and variance, thus improving prediction accuracy.
- Fine-tuning these parameters is essential to enhance the model's performance. It is hypothesized that an optimal configuration of these parameters will lead to better model generalization and improved predictive power on unseen data

6.2.2. Experimental Results

- The Random Forest model with 100 trees yielded an RMSE of 0.3725, indicating weak predictive performance despite using an ensemble approach.
- The model did not identify significant feature importance, suggesting weak predictors and the need for further feature engineering and alternative models.

7. Results Analysis and Discussion

7.1. Multiple Linear Regression

- The MLR model achieved a low RMSE of 0.3717, with very low R-squared (0.00149) and adjusted R-squared (0.000143), indicating minimal explanatory power and poor predictive accuracy. The model struggles due to weak relationships between predictors and the target variable.
- The F-statistic of 1.11 and p-value of 0.275 suggest that the overall model does not significantly improve predictions compared to a model with no predictors, confirming the model's inability to accurately predict the target.

7.2. Random Forest Regression

- The Random Forest model showed an RMSE of 0.3725, slightly worse than MLR, indicating weak predictive performance despite using an ensemble method (100 trees). This suggests that the data does not have strong signals for accurate predictions..
- Despite the use of Random Forest, the model failed to identify meaningful feature importance, highlighting that the predictors in the dataset do not contribute significantly to the target variable, suggesting the need for further feature engineering.
- In both training and when generalizing the model to the final test set, there is a large discrepancy between the resubstitution error and the validation and out-of-bag errors.

8. Future Directions

8.1. Lessons Learned

In comparison to a Both Random Forest and Multiple Linear Regression showed weak predictive performance with minimal differences in RMSE, reflecting their shared inability to address the dataset's limitations. While Random Forest, designed for complex relationships, failed to identify significant features, MLR struggled with weak linear correlations. These results highlight the need for robust feature engineering, as the dataset lacks refined predictors. Furthermore, advanced modeling techniques alone are insufficient; exploring methods like neural networks or support vector machines, paired with improved data preprocessing, is essential for better accuracy.

8.2. Future Directions

- One key future direction involves refining the dataset by conducting thorough feature selection, transformation, and possibly incorporating domain knowledge. Enhanced feature engineering could lead to more meaningful predictors, improving the performance of both linear and ensemble models.
- Given the limited success of the current models, exploring more advanced machine learning techniques such as deep learning, gradient boosting machines, or hybrid models might be a promising approach. These methods can better capture complex patterns and interactions