Franchises Database Analysis

Pedro Bernardo Jr (2126418)

Felipe Bezerra (2017557)

Filipe Romano (2117537)

Mateo Huascar (2105799)

Karina Mejía (2125224)

University Canada West

Professor Rushdi Alsaleh

May 20th, 2023

Introduction

The following paper addresses the questions about how the franchise business is performing; the database contains a conjunction of 100 franchises and is split into six main categories, net profit, counter sales, drive-through sales number of customers, business type and location. The data of the first three columns are measured in (million $), and the number of customers was taken daily; the business type can be either a pizza store, burger store or Café. Finally, the Location is where the franchise is located, and it can be Vancouver or Richmond.

Analysis

a) Decision tree model for net profit prediction and assessment of the accuracy

We have a Mean Absolute Error (MAE) for the decision tree model, which measures the average absolute difference between predicted and actual values. An MAE of 30,000 indicates a low average difference between predicted and actual values. On the other hand, Mean Squared Error (MSE) measures the average squared difference between predicted and actual values. A low MSE suggests a slight average squared difference but does not determine the presence of outliers. R-squared, which measures the proportion of variance explained by the features used in the model, is 0.97, indicating that 97% of the variability in the target variable can be explained. However, it is essential to consider other factors like overfitting and dataset characteristics before concluding the model's performance.

R-squared: This metric measures the proportion of variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with 1 indicating a perfect fit. An R-squared of 0.97 means that 97% of the variability in the target variable can be explained by the features used in the model, which is an excellent result.

These metrics suggest that the model has a robust predictive performance. However, it is essential to consider other factors like overfitting and the nature of the dataset before concluding that the model is excellent. It is adequate to cross-validate your results using techniques like k-fold cross-validation to assess the model's performance on different subsets of the data. This will give you a more reliable estimate of the model's generalization ability to new data.

A MAPE value of 1.76 means that, on average, the model's predictions are about 1.76% off the actual values. For example, if a particular franchise's actual net profit were $100,000, the model would, on average, predict it to be approximately $98,240 (1.76% lower) or $101,760 (1.76% higher). The MAPE represents the average of the absolute percentage differences between the actual and predicted values, providing a straightforward way to understand the model's prediction accuracy in percentage error. Considering such a low MAPE, the model has excellent accuracy.

Therefore, the model has high accuracy and strong predictive performance. The predicted values are very close to the actual values, providing insights into believing they have low variance and are relatively precise. A high R-squared value indicates a good fit of the model to the data.

b) Simulation of the decision tree, visualization, and interpretation of descriptive characteristics impact

The feature importance provides insight into the relative importance of the descriptive features in the decision tree model. According to the feature importance, the most important feature is "Business Type_Burger store," with an essential score of 0.423, followed by "Drive-through Sales" (0.247) and "Counter Sales" (0.225). This suggests that "Business Type_Burger store" has the most decisive impact on the model's predictions, followed by the other two features.

The permutation importance further supports these findings. The ranking of features based on permutation importance aligns with the feature importance, confirming the significance of

"Business Type_Burger store" (1.069), "Drive-through Sales" (0.478), and "Counter Sales" (0.399). These results reinforce the importance of these features in the model. The role of each feature will be explained separately in the following lines.

Regarding the business type, we can argue that The "Business Type_Café" feature also plays a role in the decision tree splits. When the business type is classified as a café and other conditions are met, the predicted net profit ranges from 0.937 to 1.655. This indicates that being a café can have a moderate positive impact on net profit. Meanwhile "Business Type_Pizza Store" feature is not involved in any splits in the decision tree, indicating that it may have less influence on the predicted net profit than burger stores and cafés.

In the case of sales performance, the "Counter Sales" and "Drive-through Sales" features are important in distinguishing different subsets of data within specific business types. The decision tree splits on these features to capture variations in net profit. However, "Counter Sales" and "Drive-through Sales" values determine the predicted net profit in different ranges. For example, lower "Counter Sales" values (e.g., <= 4.6mi) tend to be associated with lower predicted net profit values (e.g., 1.325mi). In contrast, higher "Drive-through Sales" values (e.g., > 2.95mi) can result in higher predicted net profit values (e.g., 1.861mi).

For the other descriptive features, the "number of customers" feature also contributes to the decision tree splits, indicating its relevance in predicting net profit. However, its impact is less prominent than the business type and sales performance features. The "Location_Richmond" and "Location_Vancouver" features do not appear in any splits, suggesting that they may have a limited impact on the predicted net profit in the given decision tree.

Based on the decision tree structure, the business type (specifically burger stores and cafés) and sales performance (counter sales and drive-through sales) have the most significant impact on

the predicted net profit. Other descriptive features, such as the number of customers and location, may have a lesser but noticeable influence.

c) Development of a random forest model for the net profit prediction

We have a Mean Absolute Error (MAE) of 79,400 for the random forest model, which suggests the predictions are quite close to the actual values, the same as the decision tree model developed before.

In the case of the R-squared, we got 0.96, which means that 96% of the variability in the target variable can be explained by the features used in the model, which is also a great result; however, the decision tree model gave us a better result. As we saw in the decision tree model, we need to consider other factors like overfitting and the nature of the dataset before concluding that the model is indeed excellent, so we will cross-validate the results again with the k-fold cross-validation to assess the model's performance on different subsets of the data. Furthermore, the model appears to have high accuracy and strong predictive performance.

The predicted values are very close to the actual values, providing insights into believing they have low variance and are relatively precise. A high R-squared value indicates a good fit of the model to the data.

This model's MAPE is 6.61, similar to Decision Tree but slightly less accurate. We have a deviation of nearly 6 percent above and below the predicted value. When comparing our metrics (r-squares, MAE and MSE), especially MAPE, we can tell that Decision Tree suits this dataset.

d) Rationalization of the structure of the random forest model

In defining the structure of the Random Forest model, we used a parameter grid (param_grid) to assess three different parameters: the number of estimators, the maximum number

of features, and the maximum depth. This exercise aims to optimize the model by maximizing accuracy and preventing underfitting or overfitting.

The Random Forest model does not necessarily begin with a predetermined feature; instead, it could initiate with any feature, including the number of customers, service type, business type, or even the city where the franchise is located. It then leverages randomness to build multiple decision trees, starting with different features.

The model determines the discrepancy between the training dataset and the predicted dataset, utilizing the specified maximum number of estimators, features, and maximum depth. The primary goal is to minimize this discrepancy or error while balancing the complexity of the model to avoid overfitting or underfitting.

It's crucial to remember that while this randomized approach of the Random Forest model brings the benefit of ensembling and robustness to outliers, it doesn't guarantee that the most significant features will always be at the top of the trees, unlike in the Decision Tree model where features are strictly ranked based on their importance.

e) Simulation of random forest parameters, visualization, and interpretation of descriptive characteristics impact

In the Random Forest model, we observed that the ranking of feature importance aligns with that of the Decision Tree model. The "Business Type_Burger store" continues to dominate with a score of 0.352, followed by "Drive-through Sales" (0.262) and "Counter Sales" (0.202). These results correlate with those from the Decision Tree model, indicating the robustness of these features' influence across both models. The permutation importance further underscores the significance of "Business Type_Burger store" (0.501), "Drive-through Sales" (0.425), and "Counter Sales" (0.316).

Looking at the categories individually, "Business Type_Café" and "Business Type_Pizza Store" contributed to the Random Forest model with scores of 0.112 and 0.045, respectively. Although these scores are less than "Business Type_Burger store," their non-negligible contributions suggest that different business types impact net profit differently.

In terms of sales performance, "Drive-through Sales" (0.262) seem to have a more substantial effect on franchise profitability than "Counter Sales" (0.202), though both are among the top three significant features. The remaining features had less than 0.03 significance, with the location being less influential than the number of customers.

Comparative analysis of the Decision Tree and Random Forest models offers valuable insights into their respective predictive performance for franchise net profit. The Decision Tree model outperformed the Random Forest model regarding Mean Absolute Error (MAE) and R-squared. Its lower MAE (0.03) suggests that its predictions were, on average, closer to the actual net profit values than those of the Random Forest model (MAE 0.08). The Decision Tree model's higher R-squared value (0.97) indicates it explained 97% of the net profit variance using the selected features.

Despite having identical Mean Squared Error (MSE) values of 0.01, the Decision Tree model demonstrated more robust accuracy and explained variance. This difference suggests that the Decision Tree model's more straightforward structure better captured the underlying data relationships, leading to improved predictions. Hence, the Decision Tree model might be more suitable for predicting franchise net profit in this specific case.

The Decision Tree model's structure, which splits the dataset based on the most informative features like "Business Type," "Drive-through Sales," and "Counter Sales," was pivotal in its predictive superiority. Its fewer splits helped avoid overfitting, leading to a better generalization of unseen data. The feature importance analysis reveals that "Business Type," mainly "Business

Type_Burger store," had the most significant impact on net profit predictions, aligning with our domain knowledge that business type significantly influences profitability. Moreover, "Drive-through Sales" and "Counter Sales" offered valuable insights into sales performance, enabling the model to capture net profit variations based on these metrics.

In this particular case, the Random Forest model's performance was slightly lower, which could be due to various factors such as the specific data distribution or parameter settings. While Random Forests harness the collective wisdom of multiple trees to make predictions, in some cases, they might not outperform a well-tuned decision tree.

In the Random Forest model, we observed that the ranking of feature importance aligns with that of the Decision Tree model. The "Business Type_Burger store" continues to dominate with a score of 0.352, followed by "Drive-through Sales" (0.262) and "Counter Sales" (0.202). These results correlate with those from the Decision Tree model, indicating the robustness of these features' influence across both models. The permutation importance further underscores the significance of "Business Type_Burger store" (0.501), "Drive-through Sales" (0.425), and "Counter Sales" (0.316).

Looking at the categories individually, "Business Type_Café" and "Business Type_Pizza Store" contributed to the Random Forest model with scores of 0.112 and 0.045, respectively. Although these scores are less than "Business Type_Burger store," their non-negligible contributions suggest that different business types impact net profit differently.

In terms of sales performance, "Drive-through Sales" (0.262) seem to have a more substantial effect on franchise profitability than "Counter Sales" (0.202), though both are among the top three significant features. The remaining features had less than 0.03 significance, with the location being less influential than the number of customers.

Comparative analysis of the Decision Tree and Random Forest models offers valuable insights into their respective predictive performance for franchise net profit. The Decision Tree model outperformed the Random Forest model regarding Mean Absolute Error (MAE) and R-squared. Its lower MAE (0.03) suggests that its predictions were, on average, closer to the actual net profit values than those of the Random Forest model (MAE 0.08). The Decision Tree model's higher R-squared value (0.97) indicates it explained 97% of the net profit variance using the selected features.

Despite having identical Mean Squared Error (MSE) values of 0.01, the Decision Tree model demonstrated stronger accuracy and explained variance. This difference suggests that the Decision Tree model's more straightforward structure better captured the underlying data relationships, leading to improved predictions. Hence, the Decision Tree model might be more suitable for predicting franchise net profit in this specific case.

The Decision Tree model's structure, which splits the dataset based on the most informative features like "Business Type," "Drive-through Sales," and "Counter Sales," was pivotal in its predictive superiority. Its fewer splits helped avoid overfitting, leading to a better generalization of unseen data. The feature importance analysis reveals that "Business Type," mainly "Business Type_Burger store," had the most significant impact on net profit predictions, aligning with our domain knowledge that business type significantly influences profitability. Moreover, "Drive-through Sales" and "Counter Sales" offered valuable insights into sales performance, enabling the model to capture net profit variations based on these metrics.

In this particular case, the Random Forest model's performance was slightly lower, which could be due to various factors such as the specific data distribution or parameter settings. While Random Forests harness the collective wisdom of multiple trees to make predictions, in some cases, they might not outperform a well-tuned decision tree.

In summary, the chosen structure of the Decision Tree model, informed by the most significant features and minimal complexity, contributed to its superior predictive performance. It effectively captured the relationship between the input features and net profit, making more accurate predictions. Although still performing reasonably well, the Random Forest model was slightly outperformed by the Decision Tree model in this particular instance.

f) Forecast of the net profit, when the counter sales are equal to 500,000 $, the drive-through sales are equal to 700,000 $, and the business model is a Pizza store in Richmond, using the decision tree and the random forest model—comments about the results.

When running the suggested input in both models, we have different predictions output:

- Decision Tree predicted net profit: $200,000.00.
- Random Forest predicted net profit: $326,000.00.

This suggests that the Random Forest model predicts a more extensive net profit than the Decision Tree model. Although this might be precisely what a business owner wants to hear, it may reflect a different prediction. Because Decision Tree had lower MAPE, MAE and MSE and a higher percentage of R-squared, the team's understanding is that the forecast of such net profit would be closers to 200k rather than 326k.

g) Roles of "max_feature" and "n_estimators" in a random forest model.

The "max_features" parameter in a random forest model determines the number of features to consider when looking for the best split at each node. It helps control the randomness of feature selection, thereby impacting the diversity and performance of the model. A smaller value can lead

to increased variety among trees. Still, it may risk reducing accuracy, while a larger value can result in more homogeneity among trees, possibly improving model performance but at the cost of decreased model diversity.

The "n_estimators" parameter specifies the number of decision trees to be included in the random forest. Predictions are made by averaging or taking the majority vote among all the trees, each constructed independently using a subset of the data. Increasing the number of estimators can improve the model's performance by decreasing overfitting risk and enhancing prediction stability. However, there's a trade-off, as more trees increase computational requirements and beyond a certain point, additional trees may not significantly improve performance.

h) Assumptions regarding the limitations of both models.

Models like random forests make certain assumptions and carry limitations, which may vary based on the specifics of the problem and available data. Here are some common assumptions and limitations:

Assumptions: Models often assume independence among observations. While many models assume linearity among their variables, random forests can handle non-linear interactions. Models also assume that the included features are relevant and have predictive power over the target variable.

Limitations: Overfitting can occur in a random forest if the model is too complex or the number of trees is too large. Regularization methods can help reduce this. Random forests, being complex models, may be harder to interpret than simpler models like linear regression. Identifying individual feature importance can be challenging.

Random forests require substantial data for practical training and validation, and poor results can stem from insufficient or non-representative data. They can be sensitive to outliers, which may impact model performance.

Outliers also have an integral whole on how well (or not) Decision Tree or Random Forest models perform. We saw in item f that the model couldn't predict well a low number of sales because both training and testing data were only exposed to bigger magnitudes numbers (e.g., average net profit is 1.45 million). Therefore, when predicting 1.2 million of total sales, the model wouldn't extrapolate much from the dataset maximums and minimums.

The computational complexity of random forests increases with the size of the dataset or the number of features, making them potentially unsuitable for vast datasets. Evaluating the specific context and data before applying any modelling approach is always prudent, as assumptions and limitations can differ.

Appendix

See the ZIP file for further calculations, visualizations, and coding information.