Revenue forecast as the foundation for dynamic pricing

Business Problem:

Real-time automated price generation moves into focus for online retailers. Every day, category management faces the challenge of adapting thousands of products to changing market factors. We can easily predict the revenue when the prices are static but it is challenging to predict the revenue when the prices are generated dynamically. This happens completely automatically when using an intelligent pricing tool. Prices reflect customer appreciation and lead to increased sales.

To help better understand the influence of Dynamic pricing on revenue and make more accurate revenue predictions, we will develop a statistical model for predicting future revenue using historical anonymised transaction data of a real mail-order pharmacy found online. The special feature of this dataset is the product prices were dynamically generated automatically.

Dataset:

This dataset has around 660K rows with 21 Attributes. It contains the details of a 30 day transaction of a pharmacy. It has information about products being sold and their information like competitor price, revenue, manufacturer etc.

Approach:

1. Data Preprocessing

All features were analyzed to identify potential outliers and relationships among them. Further, to hold significant information, missing values have been handled using KNN Imputer and other strategies.

2. Data Transformation

Following is the pipeline we have created to apply encoding and scaling to the attributes based on its characteristics and values. We did frequency encoding (which has more than 50 categories in order to get rid of the curse of dimensionality) and one hot encoding for the categorical columns.

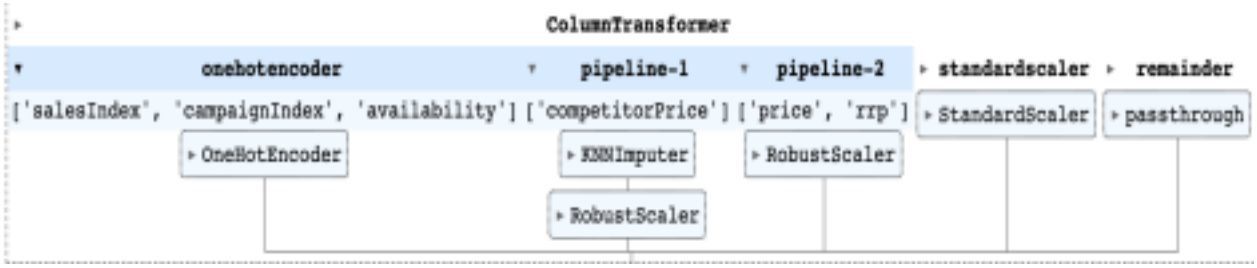

Fig (a): Pipeline Transformer

For features having outliers, robust scaling was applied and for other features, standard scaler has been used after studying them.

3. Data Modeling

There are many different statistical models that we can use to predict revenue when prices are dynamic. Ultimately, the best model for predicting revenue when prices are dynamic will depend on the characteristics of the data and the specific goals of the prediction. It is necessary to try out various models and compare their performance in order to find the one that works best. Hence, we have applied the following regression models which we felt suitable for this.

- Linear Regression (Linear, Lasso and Ridge)
- Decision tree models (Decision Tree, Random Forest and XGBoost regressors) -

SVM Regression

- KNN Regression

We have obtained above results in which tree models such as Random Forest and XGBoost regressors worked better among all.

```
{'lr': 0.47663633722933446,
 'ridge': 0.4766358153882839,
 'lasso': 0.4312964771416225,
 'knn': 0.708078215486331,
 'dtree': 0.699633862543035,
 'adaBoost': 0.3862489665568953,
 'randomforest': 0.7465414478676241,
 'xgb': 0.7659135582554805}
```

Fig (b) : R2 score obtained for models

As we have got around 75% of accuracy for XGB and Random Forest, we have applied hyperparameter tuning for both the models. We found that the XGB regressor was the most effective model for predicting revenue in our case as it performed well on both training and validation data.

- Neural Network

To check how a Neural Network model works on our data, we trained it with a basic NN architecture. Even though it is a known fact that the neural network does not work as well as it works with images compared to the tabular data, our model was able to get a decent R2 score. Training it with an improvised architecture having more epochs and computational power, the model will likely outperform the basic architecture.

4. Predictive Validation

We have below scores on the test data which concludes that the model shows high predictive validity as they are comparable to the score on training data.

R2 score for XGB and Random Forest : 76%

R2 score for Neural Network: 74%

Findings:

The below curve depicts that the error is decreasing and both the training and validation curves are converging. Therefore, the model is neither overfit nor underfit and considered as the best fit.

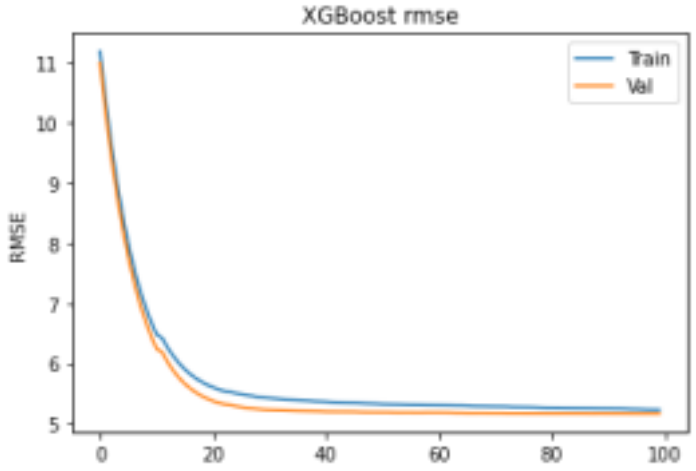

Fig (c): Learning Curve for XGB Model using RMSE metric

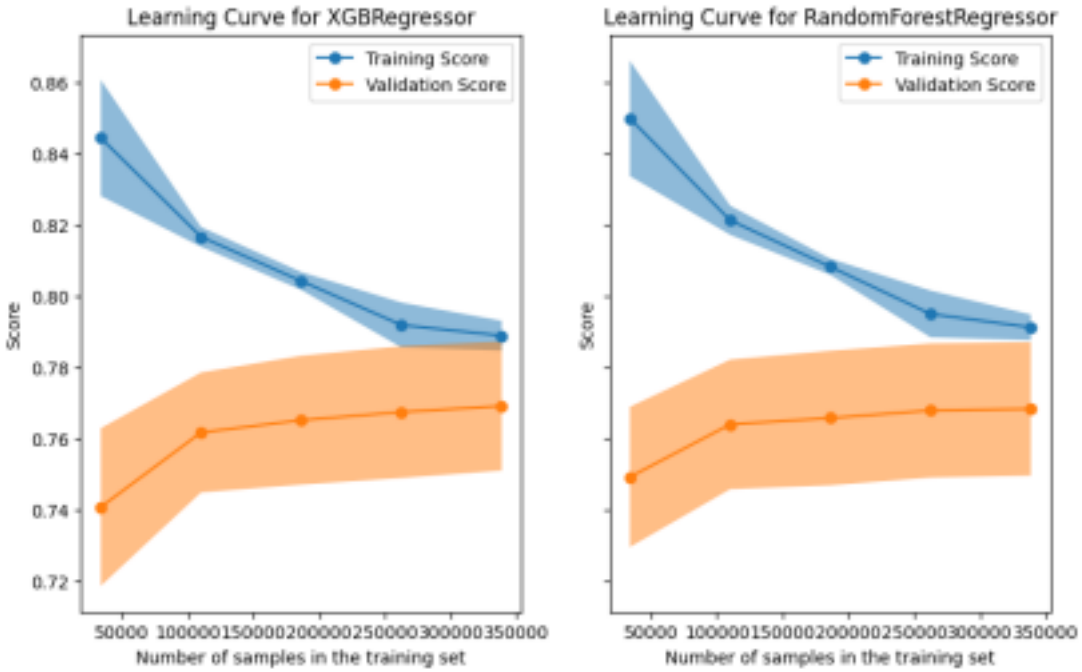

Fig (d): Learning Curve for XGB and RF Models using R2 score metric

Above Learning curves describes that:

- The training and validation curves are not yet converged which means the model is neither overfit nor underfit.
- Even though the R2 value decreases for the training data, the test score is increasing with the number of data points. This suggests that having more data points would improve the score on test data.

Conclusion:

After exploring several different modeling approaches, we found that XGBoost and Random Forest regression are the most effective models for predicting revenue. Our findings suggest that businesses can use statistical modeling to make more accurate revenue predictions in dynamic price environments and better understand the factors that influence prices. By analyzing past data and making assumptions about future price changes, businesses can make informed decisions about pricing strategies and optimize their revenue streams.

Futurescope:

Training the model with huge data would give more scope to accurately predict the revenue. In addition, feature engineering would help improve the accuracy of the model. Also, to increase the prediction accuracy of a neural network model, we can use a sophisticated architecture with rigorous training. Furthermore, we can try using other state-of-the-art models like Tab Transformer or other self attention mechanisms to experiment how it works.

References:

1. https://scikit-learn.org/stable/auto_examples/model_selection/plot_learning_curve.html#sphx-glr-auto-examples-model-selection-plot-learning-curve-py
2. https://towardsdatascience.com/xgboost-fine-tune-and-optimize-your-model-23d996fab663
3. https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd7