Xin Tang
Predictive Analytics
DSC630-T302 (2247-1)
Dr. Andrew Hua

Week 5

Predictive analytics case study

Used car price prediction case

Introduction:

I do not have any predictive case in my daily life or work. So, I turned to the internet for PA case examples. Most are very general case briefings with no detailed code, nor coverage of the whole process. I ended up on the coding cases published in Kaggle. I picked a car selling price data coding case and provided my own case study.

Problem to be solved:

- Using the information collected from past used car sales, i.e.: car model name. model year. KM driven, transmission type etc. the author tries to build a machine learning model to predict the used car selling price.

Why was this problem important to solve?

- (Athira, 2023): I was interested in cars always and wanted to know the factors affecting prices.

- “The information problems that exist in the used car market were potentially present to some degree in all markets “(George A. Akerlof, 2001). This statement from a Nobel prize winner is evidence that buying a used car is a difficult task. Most buyers in the used car market have no clear reference of the reasonable price, nor what car is in higher demand. This case is to use the historic car selling data to predict car price. It intended to build a few predictive models which can help to find out the proper values.

How was the data acquired?

- The data is from Kaggle website, but the origin and author are not clearly mentioned.
- The data consists of 8 variables and information from 300 cases of car sales.

Methods and Results

What steps were taken to prepare the data?

- First, basic understanding of the data. Including check data shape, variable types and unique values.
- Second, data visualization exposes the distribution of each variable.
- Variable type adjustment to fit for data modeling.
- Heatmap to study correlation between variables.
- Data cleaning: such as removing duplicate and missing values.
- Feature encoding to treat nominal data and ordinal data

- Outlier detection has also been performed to replace the selling price outliner with median value.
- Last feature selection and scaling are both performed

How was this problem solved?

- The whole process mainly followed the CRISP process. It went through data preparation, model training, model optimization, evaluation and selection process.
- After model buildings, various models were evaluated.
- Based on evaluation, a model was suggested.

What modeling techniques were used?

- This is a regression problem, several regression modeling techniques were used: Random Forest regressor, SVR, linear Regression, XG Boost regressor, CatBoost regressor and LGBM regressor.
- Then hyper-parameter tunning was performed on LGBM model, XG boost model, Random Forest model and Decision tree model.

Why did the team choose the methods/models they did?

- Since the problem is to predict a numerical/continuous value, it is a regression problem. Various regressors models are used.
- Since this is an exercise coding, so author picked as many models as possible to pick and choose.

What metrics were used to evaluate the results? Why were these metrics chosen?

- Several metrics were used to evaluate the models. Like accuracy, R2 score, MAE, MSE and RMSE values.
- The R2 on both training and test set has also been evaluated to find out if models are overfitting or underfitting.
- These are the most used metrics and apply to almost all the regression models. They also give quite straight result of the model performance.

Conclusion

How were the results or model implemented?

- The code did not give the information if the model was implemented.

What were the actionable consequences of the case study?

- Since this PA case covers a very common business application and apply to both individual and enterprise. After the model has been built and selected, it could be deployed for actual use. Any individual car buyer could apply this model to guide their car buying process; for enterprise like car dealers, used car collectors etc., they can use it to evaluate their car trading and price the car appropriately.

What did the team learn from the case study?

- The author could conclude that a stacking model is the best for this PA case.

How should or would the team approach the problem differently in the future?

- For this case, I did not see any business expert involvement, they could provide valuable input on if the data is accurate, any other variables are important and needed to be included in the analysis, also they could help to evaluate the prediction result from the business use angles.

Resources:

1. Case study dataset: https://www.kaggle.com/datasets/athirags/car-data/data
2. Code: (Mehedi, 2024): https://www.kaggle.com/code/mehedithedreamer/predicting-car-selling-price
3. Scott A. Wolla (2016) Why Is It So Difficult to Buy a High-Quality Used Car? https://files.stlouisfed.org/research/publications/page1-econ/2016-09-01/why-is-it-so-difficult-to-buy-a-high-quality-used-car_SE.pdf