Retail Sale Prediction

Data science trainee

AlmaBetter, Bangalore

SAI NARASIMHA CHARAN THULASI

Abstract:

The Retail Sales Prediction dataset is provided with the two datasets first one is the rossman dataset and second one is the store dataset both of the datasets contains different information and they are both equally important datasets in respect to our project.Sales are the lifeblood of a business. It's what helps you pay employees, cover operating expenses, buy more inventory, market new products and attract more investors. Sales forecasting is a crucial part of the financial planning of a business. It's a self-assessment tool that uses past and current sales statistics to intelligently predict future performance.

With an accurate sales forecast in hand, you can plan for the future. If your sales forecast says that during December you make 30 percent of your yearly sales, then you need to ramp up manufacturing in September to prepare for the rush. It might also be smart to invest in more seasonal salespeople and start a targeted marketing campaign right after Thanksgiving. One simple sales forecast can inform every other aspect of your business.

Sales forecasts are also an important part of starting a new business. Almost all new businesses need loans or start-up capital to purchase everything necessary to get off the ground: office space, equipment, inventory, employee salaries and marketing. You can't just walk into a bank with a bright idea and lots of enthusiasm. You need to show them numbers that prove your business is viable. In other words, you need a business plan.

A central part of that business plan will be the sales forecast. Since you won't have any past sales numbers to work with, you'll have to do research about related businesses that operate in the same geographical market with a similar customer base. You'll have to make concessions for the difficulty of starting from scratch, meaning that the first few months will be lean. Then you'll need to convince the bank that your business has fresh ideas that will eventually outsell the competition. All of these ideas need to be expressed as numbers -- losses, profits and sales forecasts that the bank can easily understand.

2. Introduction

The dataset contains the data of the two Datasets that is one of the Rossman dataset and second one is the store dataset.If we combine both of the datasets then we will get about 19 columns and 1017209 rows which contains different kind information which is helpful to take better descisions using data.The project mainly focuses on the sales forcasting for given data of stores and deploying various models.In this project I have used two models first one is the Linear

Regression model And second one is the Lasso regression model

3. Problem Statement

Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Rossmann store managers are tasked with predicting their daily sales for up to si x weeks in advance. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality. With thousands of individual managers predictin g sales based on their unique circumstances, the accuracy of results can be quite varied.

You are provided with historical sales data for 1,115 Rossmann stores. The task is to forecast the "Sales" column for the test set. Note that some stores in the dataset were temporarily closed for refurbishment

4. Data summary

The dataset contains two data sets that one of the rossman dataset and second one is the store datasets if we combine both of the datasets there is 19 columns and 1017209 rows the column contains different entries with different data such an Id that represents a (Store, Date) duple within the test setStore a unique Id for each store Sales the turnover for any given day (this is what you are predicting) Customers the number of customers on a given day, Open an indicator for whether the store was open: 0 = closed, 1 = open,StateHoliday indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None SchoolHoliday indicates if the (Store, Date) was affected by the closure of public schools, StoreType differentiates between 4 different store models: a, b, c, d Assortment describes an assortment level: a = basic, b = extra, c = extended CompetitionDistance distance in meter s to the nearest competitor store CompetitionOpenSince[Month/Year] gives the approximate year and month of the time the nearest competitor was openedPromo indicates whether a store is running a promo on that day Promo2 Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is par ticipating Promo2Since[Year/Week] describes the year and calendar week w hen the store started participating in Promo2PromoInterval - describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew. E.g. "Feb,May, Aug,Nov" means each round starts in February, May, August, November of any given year for that store

5. Steps involved

Various steps involved in this process to analyze the data **Importing required libraries**

Mounting drive

Data cleaning

EDA(Exploratory Data Analysis)

HypothesisTesting

Outlier detection/handling

Feature Engineering

Model deployment

i)Importing Required libraries

The libraries which are used in this project are numpy and pandas which is imported first and for the data visualization we have used matplotlib as well as seaborn libraries.In this project we have created different pie charts,bar charts,heat maps etc by using the both of libraries.Numpy and Pandas are both very useful while Performing

Exploratory Data Analysis.

ii)Mounting Drive

Next step is related to mounting the csv file,After mounting the csv file to colab notebook we have able read the all data and get info regarding data that means how much rows and columns etc.

iii)Data Cleaning

1) The first step involves the checking and removing of duplicated rows in this step we have used value count method for duplicate rows and removed the values.

The next step involves the Handling of missing values in this step we handled values using null function after getting result we found that in Rossman dataset there is no null values but in the store dataset 6 columns contains the missing values in the first step we have done competition distance column filled with the median value because the distribution I found that is right skewed and some columns which not have any use they are dropped

iv)Exploratory Data Analysis

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.I have done multivariate analysis and found some very interesting facts about data

sales are highly correlated with customers,there were more sales on Monday,probably because shops are closed on Sundays,it could be seen that the promo leads to more sales,more stores open on school holidays than on state holiday and hence had more sales than state holidays,on an average store type B had the highest sales,Highest average sales were seen with Assortment levels-b which is ‘extra’,82.1% sales are not affected and only 17.9% sales is affected because of school holiday etc.

HypothesisTesting

It is also one of the important steo in the project Hypothesis testing is the process used to evaluate the strength of evidence from the sample and provides a framework for making determinations related to the population, ie, it provides a method for understanding how reliably one can extrapolate observed findings in a sample under study to the larger population

So I have also done hypothesis below

Hypothesis: Stores located closer to com petition have significantly lower sales th an stores located further away. Null hypothesis: There is no significant difference in sales between stores locate d closer to competition and stores locate d further away.
Alternative hypothesis: Stores located cl oser to competition have significantly lo wer sales than stores located further awa y.To test this hypothesis, we can perfor m a two-sample ttest between the sales of stores located w ithin 10 kms of competition and stores l ocated further away. We can set a signifi cance level of 0.05

If the pvalue is less than the significance level (0.05), we can reject the null hypo thesis and conclude that stores located cl oser to competition have significantly lo wer sales than stores located further awa y. Otherwise, we fail to reject the null hy pothesis and conclude that there is no si gnificant difference in sales between the two groups.

Outlier detection/handling

Outliers are generally defined as samples that are exceptionally far from the mainstream of data. There is no rigid ma thematica l definition of what constitutes an outlier; determining whether or not an observation is an outlier is ultimately a subjective exercise. An outlier may also be explained as a piece of data or observation that deviates drastically from the given norm or average of the data set. An outlier may be caused simply by chance, but it may also indicate measurement error or that the given data set has a heavy-tailed distribution.
Therefore, Outlier Detection may be defined as the process of detecting and subsequently excluding outliers from a given set of data. There are no standardized Outlier identification methods as these are largely dependent upon the data set. Outlier Detection as a branch of data mining has many applications in data stream analysis.I have used Tukeys method for detection of outliers and removing them Tukey's rule says that the outliers are values more than 1.5 times the interquartile range from the quartiles — either below Q1 − 1.5IQR, or above Q3 + 1.5IQR. We will use these as part of writing a function to identify outliers according to Tukey's rule.

Feature Selection/Engineering

Feature engineering is the processof selecting, manipulating, and transforming raw data into features that can be used in supervised learning. In order to make machine learning work well on new tasks, it might be necessary to design and train better features.

First manupulation that I have done that is dropped open column then checked if any column opened with zero sales percentage of open stores with zero sales then removed it to avoid the bias then I have dropped five columns 'Sales','Store','Date','Year','StateHoliday’

Model Deployment

In models I have used two models first one is the linear regression model and second one is the lasso regression model

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.[1] This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.[2] In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models.[3] Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.
Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications.[4] This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters

Conclusion

Linear Regression Train Score and Test Score: The train sc ore of 0.788 and test score of 0.782 sugg est that the model is performing well in t erms of accuracy, with a slightly higher score on the training data compared to th e test data. In terms of business impact, t his means that the model can provide rea sonably accurate predictions for the targ et variable, which can be valuable for making decisions such as forecasting sales, estimating customer lifetime value, or predicting demand.

RMSE: The root mean squared error (RMSE) of 1174.04 suggests that, on average, the model's predictions are off by approximately 1174 units. In terms of business impact, this metric can be used to evaluate the accuracy of the model's predictions, and to identify areas where the model may need improvement. For example, if the model is being used to predict customer lifetime value, a high RMSE may indicate that the model is not accurately predicting the true lifetime value of customers, which could impact decisions related to marketing or customer acquisition.

Train RMSE and CV RMSE: The train RMSE of 1153.02 and CV RMSE of 1155.08 suggest that the model is performing well in terms of accuracy, with a slightly lower RMSE on the training data compared to the crossvalidated data. In terms of business impa ct, this means that the model can provide reasonably accurate predictions for the t arget variable, which can be valuable for making decisions such as forecasting sales, estimating customer lifetime value, or predicting demand. Additionally, the fact that the training and crossvalidated RMSE scores are similar sugg ests that the model is not overfitting to t he training data, which is important for ensuring that the model can generalize well to new, unseen data.

Lasso Regression Regression Model Score (Rsquared): The Rsquared value of 0.788 suggests that the model is able to explain approximately 7 9% of the variance in the dependent vari able using the input features. In terms of business impact, this means that the mod el is able to provide a good fit to the data, and can be used to make accurate predictions about the target variable. For exam ple, if the model is being used to predict sales, a high Rsquared value would indicate that the m odel is able to explain a large proportion of the variability in sales, and can be used to make more accurate sales forecasts.

Out of Sample Test Score: The out-ofsample test score of 0.782 indicates that the model is able to generalize well to ne w, unseen data. In terms of business imp act, this means that the model is likely to perform well when making predictions on new data, which is important for ensuring that the model can be used to make accurate predictions in realworld scenarios.

Training RMSE and Testing RMSE: The root mean squared error (RMSE) measures the average difference between the actual and predicted values of the target variable. The training RMSE of 1194.56 and testing RMSE of 1172.61 suggest that the model's predictions are, on avera ge, off by approximately 1194 and 1172 units, respectively. In terms of business i mpact, these metrics can be used to eval uate the accuracy of the model's predicti ons, and to identify areas where the mod el may need improvement. For example, if the model is being used to predict customer lifetime value, a high RMSE may indicate that the model is not accurately predicting the true lifetime value of cust omers, which could impact decisions rel ated to marketing or customer acquisition.

Training MAPE and Testing MAPE: The mean absolute percentage error (MAPE) measures the average difference between the actual and predicted values of the target variable as a percentage of the actual value. The training MAPE of 15.03 and testing MAPE of 15.60 suggest that the model's predictions are, on average, off by approximately 15% of the actual value. In terms of business impact, these metrics can be used to evaluate the accuracy of the model's predictions in a more interpretable way than RMSE, and can be used to compare the accuracy of different models. For example, if the model is being used to predict demand for a product, a high MAPE may indicate that the model is not accurately predicting the true level of demand, which could impact decisions related to production and inventory management..

References

1. **Analytics vidhya**
2. **Stack Over Flow**
3. **Wikipedia**
4. **Campus X**
5. **KDnuggets**
6. **Almabetter**