Insurance charges prediction

Dhiraj Bankar

Bellevue University

Abstract

For project 3, I choose to go with the retail dataset. I worked with Sears Holding corporation, who is pioneers of shopping malls, credit cards, layaway, home services and many more. I am here because of sears and I would like to make prediction model for retail though I don't have data from Sears but I found the data for Walmart, I want to explore a predictive model-based dataset and the current dataset provided me the necessary scope required for the project. If sears would have used the Data science and leveraged the power of analytics to save retails industry. Till last year Sears was a sinking boat like a Titanic.

Context

Historical sales data for 45 Walmart stores located in different regions are available. There are certain events and holidays which impact sales on each day. The business is facing a challenge due to unforeseen demands and runs out of stock sometimes, due to inappropriate machine learning algorithm. Walmart would like to predict the sales and demand accurately. An ideal ML algorithm will predict demand accurately and ingest factors like economic conditions including CPI, Unemployment Index, etc. The objective is to determine the factors affecting the sales and to analyze the impact of markdowns around holidays on the sales.

Table of Contents

Dataset:

Below is the URL address for the heart disease dataset.

https://www.kaggle.com/rutuspatel/retail-analysis-with-walmart-sales-data

The dataset has about 8 attributes.

- Store - the store number
- Date - the week of sales
- Weekly_Sales - sales for the given store
- Holiday_Flag - whether the week is a special holiday week 1 – Holiday week 0 – Non-holiday week
- Temperature - Temperature on the day of sale

- Fuel_Price - Cost of fuel in the region
- CPI – Prevailing consumer price index
- Unemployment - Prevailing unemployment rate

The advantage with this dataset is that it can be used to build the predictive model as the goal is to predict the sale of the day for particular Walmart store. However, the challenge with this dataset is that the amount of data is very limited and it has only 6435 rows and 8 attributes including the sales which is what I am going to predict.

Holiday Events

Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13

Labour Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13

Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13

Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

Analysis Tasks

Basic Statistics tasks

1) Which store has maximum sales

2) Which store has maximum standard deviation i.e., the sales vary a lot. Also, find out the coefficient of mean to standard deviation

3) Which store/s has good quarterly growth rate in Q3’2012

4) Some holidays have a negative impact on sales. Find out holidays which have higher sales than the mean sales in non-holiday season for all stores together

5) Provide a monthly and semester view of sales in units and give insights

Statistical Model

For Store 1 – Build prediction models to forecast demand (Linear Regression – Utilize variables like date and restructure dates as 1 for 5 Feb 2010 (starting from the earliest date in order). Hypothesize if CPI, unemployment, and fuel price have any impact on sales.) Change dates into days by creating new variable.

Select the model which gives best accuracy.

Proposed Methods

There are 3 different models that are being used on the current dataset. Given it's a prediction problem, all the chosen models were regression models. As with any prediction problem, the first model is the multiple linear regression. Followed by the multiple linear regression, the next model used here is the polynomial regression just to make sure and catch any patterns that does not perform really well on the polynomial regression. The third model that was chosen for this dataset is the random forest regression model. The rationale for this model is to make sure to try a regression model which different from traditional regression models such as multiple linear regression or the polynomial regression model.

Anticipated Issues:

As mentioned above, one of the anticipated issue that I might run into with this dataset is the dataset size. Since the data set is only of the size with 6435 rows and 8 columns, I need to be really careful so that the data is **not over fit.** However, I still decided to go with the dataset as I want to learn different ways by which I can work with a smaller dataset and still get a better accuracy. The

other issue I anticipate is to deal with different accuracies I get from models and how do I choose one over the other when these individual models give me vastly different accuracies.

References

Applied Text Analysis with Python, Benjamin Bengfort, Rebecca Bilbro & Tony Ojeda

Machine Learning with Python Cookbook, Chris Albon

https://digital.hbs.edu/platform-rctom/submission/predicting-your-casualties-how-machine-learning-is-revolutionizing-insurance-pricing-at-axa/

https://cloud.google.com/blog/products/gcp/using-machine-learning-for-insurance-pricing-optimization

https://www.variancejournal.org/articlespress/articles/Machine-Spedicato.pdf

https://www.datarobot.com/use-cases/insurance-pricing/

Dataset:

https://www.kaggle.com/bmarco/health-insurance-data

Appendix A

Categorical data understanding

https://medium.com/hugo-ferreiras-blog/dealing-with-categorical-features-in-machine-learning-1bb70f07262d

https://en.wikipedia.org/wiki/Feature_(machine_learning)

LinearRegression:

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html?highlight=linear regression#sklearn.linear_model.LinearRegression

RandomForestClassifier:

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

PolynomialFeatures:

https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html