PROPOSAL

Kartik Sharma

MSc. Data Analytics 2022-23

x21125813

Data Mining and Machine Learning – 1

Motivation

Flight Price Prediction:

World has again started gaining pace after a long time of recovering from the pandemic. With the ease in the restrictions, the offices, tourist attractions, restaurants, educational institutions have the process of reopening, the employees are returning, tourists have started boarding their flight, international student are looking forward to meeting their mates in college. In this all the most important question that comes to a travelling individual is "when and how much should I spend on a flight ticket?". The Flight dataset have details ranging from the source city, price, many other attributes from the 6 major cities in India. The motive is to find the critical features affecting the price of a flight.

Superstore Sales:

Today every entrepreneur wants to make his/her business reach new heights and every customer (product consumer) expects an ease and comfort of buying products with the highest possible discounts. A right amount of research and analysis can uncover the hidden factors that derive the sales of a particular product. It can also help the product owner to decide right time and discounts on products in order to gain more customers without hampering the profits.

Heart Disease Data:

In today's times, when every want to be successful, we tend to ignore our health. There is no meaning to success without a healthy body. 32% of global deaths in 2019 were due to cardiovascular disease [1]. Clearly there's an urgent need to put technology to its right purpose and develop a system that can predict the cardiovascular disease, before it takes a toll on an individual's life.

Research Question

Flight Price Prediction:

The objective of this research is to study on the attributes of flights in order to predict most legible prices, using the Multiple Regression and using the Logistic regression whether individual should opt for the flight or not.

Superstore Sales:

Profits and Sales are key factors in making a business grow to new heights, with the use of KNN and Multiple Regression we can predict both the factors, thus can help in determining the critical features that derive these factors, indeed helping a business to take data-driven decisions.

Heart Disease Data:

The research is to find the most critical factor affecting the cardiovascular health and to predict a cardiovascular disease before a person suffers its ill effects. Random forest or XgBoost can help in predicting the disease while Multiple Regression will determine critical features affecting cardiovascular health.

Initial Literature Review

Flight Price Prediction:

As per an estimate by FlightRadar24, there are almost an average 115,000 taking off and landing in a day. Everybody from a tourist to a student, is traveling by air, a system can be built to help these individuals, or in some case the corporates to save expenses on flight tickets. The Literatures/Papers discuss an approach of how predictive model created by applying machine learning algorithms on the historical flight data can help in achieving this objective.

By using Linear Regression algorithm [2], we can easily predict the price of a flight based on various attributes like duration of flight and season (like: - Holidays). Using Logistic regression [3], we can predict whether the flight should be booked or not, in order to minimize the expenses. The evaluation is done, and final predictor model is selected.

Superstore Sales:

A true business success is not based on how much profit it made, but its how much it learned from the past. This paper discusses the approach to handle the Big Data and using the previous sales data to predict the future sales. Linear Regression algorithm [2] is used to achieve the goal along with various metrics to evaluate the models.

Also, by using Random Forest algorithm [5] we can predict the profits for future.

Heart Disease Data:

A healthy lifestyle leads to a healthy life, but in this everchanging world, sometimes we forget to spend time looking after our health. Today a major percentage of population is suffering from cardiovascular diseases. With the power a machine learning, a solution can be developed that can help doctors and individual to detect any heart related disease based on some attributes like cholesterol and Thalassemia.

This research uses various classifiers like KNN [6] and Support Vector Classifier [7], to predict whether a patient is suffering from a cardiovascular disease based on the data collected. Vote [8] with a hybrid approach of Linear Regression and Naive Bayes [9] is used, to achieve 87.4% accuracy in predicting heart diseases.

Data Sources

Flight Price Prediction:

https://www.kaggle.com/shubhambathwal/flight-price-prediction

Flight dataset is a data collected from the domestic flights operating in 6 major cities in India by Easy My Trip. This is a secondary data consisting of 300,153 rows and 12 columns.

Superstore Sale:

https://www.kaggle.com/shekpaul/global-superstore

This dataset consists of a data collected from global superstore over a period of 4 years. The dataset is a combination of 51,000 rows and 22 columns.

Heart Disease Data:

https://www.kaggle.com/alexteboul/heart-disease-health-indicators-dataset

The dataset contains various health attributes ranging from High Blood pressure to physical activity. Though this dataset designed to predict the heart disease but the same can be used to predict high cholesterol (which is another major contributor to cardiovascular diseases). The data is combination of 253,680 rows and 22 columns.

Machine Learning Methods:

Flight Price Prediction:

1. **Logistic Regression:**
   This regression algorithm is used to predict occurrence of target variable in form of binary values. Logistic Regression works best when the complete data can be linearly divided in two halves. This algorithm is common among the data scientist and analysts to solve binary classification problems.
   For flight prediction, we can use this algorithm to decide whether an individual should take the flight or should wait for price to drop shortly.

2. **Multiple Regression:**
   This regression algorithm is used to predict value of target variable/dependent variable based on the relationship created between the independent variables. This algorithm is commonly used to predict the incomes, prices or other target attributes where these attributes are highly dependent upon the independent features in the data.
   For flight prediction, we can use this algorithm to predict the price of a flight ticket based on attributes like distance between the source and destination, holiday season and many other.

Superstore Sale:

3. **K-Nearest Neighbours:**
   KNN (K-Nearest Neighbours) is a Supervised Machine Learning Algorithm that means it will create a relationship between the independent variables to predict the target/dependent variable. Unlike the linear model, here models are based upon non-linear relationships.
   KNN can be applied to superstore data set to divide the products into categories like high and low selling products in order to predict future profits.

4. **Multiple Regression:**
   This regression algorithm is used to predict value of target variable/dependent variable based on the relationship created between the independent variables. This algorithm is commonly used to predict the incomes, prices or other target attributes where these attributes are highly dependent upon the independent features in the data.
   For Super Store dataset, we can use this algorithm to predict the future sales based on attributes like holiday season, weather and many other factors.

Heart Disease Data:

5. **Random Forest:**
   A supervised machine learning algorithm commonly used for solving the classification and regression problems. This Ensemble technique [5] is the core of this algorithm, hence while being highly dependent on hardware capabilities, the results from this algorithm are highly accurate.
   This algorithm handles the categorical data in an effective manner, thus while creating a model to predict heart disease, where results are critical, this algorithm will be very effective and trustworthy.

6. **Multiple Regression:**
   This regression algorithm is used to predict value of target variable/dependent variable based on the relationship created between the independent variables. This algorithm is commonly used to predict the incomes, prices or other target attributes where these attributes are highly dependent upon the independent features in the data.
   For Heart Disease dataset, we can use this algorithm to predict critical features that contribute to cardiovascular health.

Evaluation Methods:

1. **Adjusted R2(Goodness of Fit):**
   This is used as preferred method of analysing the accuracy of a regression model in case of multiple regression. This performs better than the R2[14] because unlike R2 it does not assume that variance increases with increase in the features in model.

2. **Root Mean Squared Log Error (RMSLE):**
   RMSLE [12] takes the log of Root Mean Squared Error [13] thus reduces the scale of error. The output is number based on which one can decide how accurate will be the model.

3. **Confusion Matrix:**
   In Classification Problem, the values of target variable can be of 2 or more classes, in this case confusion matrix becomes very effective in determining the performance of the model obtained.

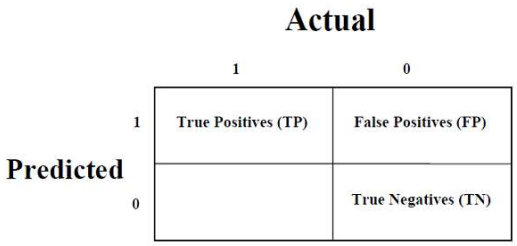

Figure 1 – Confusion Matrix [10]

4. **Recall or Sensitivity:**
   It is the ratio of True Positives to sum of True Positives and False Negatives. The idea behind this evaluation is to get the accurate number of how many True Positive outcomes, the model was able to predict.

Bibliography

[1] Cardiovascular diseases (CVDs)
https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)#:~:text=Cardiovascular%20diseases%20(CVDs)%20are%20the,%2D%20and%20middle%2Dincome%20countries.

[2] Linear Regression
By Rohith Gandhi
https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a

[3] Logistic Regression
By Jason Brownlee
https://machinelearningmastery.com/logistic-regression-for-machine-learning/

[4] Multiple Regression
By Adam Hayes
https://www.investopedia.com/terms/m/mlr.asp

[5] Random Forest
By Sruthi E R
https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[6] KNN
By Joos Korstanje
https://realpython.com/knn-python/

[7] Support Vector Classifier
By Bruno Stecanella
https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/

[8] Vote
By Aashish Nair
https://towardsdatascience.com/combine-your-machine-learning-models-with-voting-fa1b42790d84

[9] Naive Bayes
By Sunil
https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

[10] Confusion Matrix
By Tutorials Point®
https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_algorithms_performance_metrics.htm

[11] Recall or Sensitivity
By Tutorials Point®
https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_algorithms_performance_metrics.htm

[12] Root Mean Squared Log Error (RMSLE)
By Raghav Agrawal
https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/

[13] Root Mean Squared Log Error (RMSLE)
By Raghav Agrawal
https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/

[14]R2
By Raghav Agrawal
https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/