Data Mining And Machine Learning

***Abstract*—In this digital era, each business is exploded with a huge amount of digital data and it has become very difficult to draw any insights from that worthy data. While difficult, to carry out and make predictions becomes a major key to success in any business. One of the important factors to increase sales, the pricing should be deeply read and invested in by any organization. To facilitate all these into a business, predictive and analyzing machine learning models come into the picture. This project aims to use 5 machine learning techniques on 3 datasets to create a price prediction model for Airbnb and Second-hand cars and a sales model for Iowa Liquor. The models used in this study are K-Nearest Neighbors, Random Forest, Support Vector Regressor, Linear Regression, and Xgboost regressor. The results show the performance of the models and the leading factors affecting sales of the liquor across the united states and pricing of the rental property and second-hand cars. This kind of approach can be useful to predict and forecast the different features important in a business and work on those to successfully run a company/business.**

I. INTRODUCTION

People are using the internet today to do almost everything in their day-to-day life. This list includes some common things like buying and selling, finding restaurants nearby, and also renting a property. As per the Airbnb data, there are 2.9 million hosts and more than 7 million listings on Airbnb worldwide in 2020. Likewise, the second-hand automotive industry was valued at USD 1,332.3 billion in 2019 and is forecast to hit USD 1,402.0 billion in 2020. These numbers are increasing and this trend is likely to continue and increase in the future. As the sales and pricing are highly correlated, this adds additional significance to the price forecast issue. To tackle this, machine learning models are used in predicting the price of their listings. It helps in developing and improving different business strategies by increasing marketplace knowledge.

In terms of sales and price forecasting, it is important to discover insights and complex relationships between different features of the data collected. Machine learning models can allow us to derive these insights and interactions and therefore enable us to forecast a fair price that is often appealing to consumers, resulting in a rise in revenue. It can also be used to estimate potential revenue in line with existing sales. A series of machine learning models have been used for prediction purposes in this research. The models used are K-Nearest Neighbors, Random Forest, Support Vector Regressor, Linear Regression, and Xgboost regressor. These methodologies were extended to three separate datasets. Models are programmed using the R programming language.

Support Vector Regression model is applied to the first dataset i.e Iowa liquor to use the full sales data for the previous year and to create a model that could estimate the overall future sales.

In the second dataset i.e Used cars listings from Craigslist, KNN, and Random Forset for the regression model is applied to predict the price of the used vehicle based on its specifications. Mainly, Correlating the price and their specifications.

Linear and XGBoost regression model is applied to the third data set which is about property listings on Airbnb. It aims to predict the price of an Airbnb per night depending upon factors like room type, cancellation policy, property type, and Number of bedrooms, etc.

This paper follows the following sequence: Section 1 offers a brief introduction to the forecasting of sales and pricing and the methodologies used. Section 2 points out the relevant analysis that has been carried out. Section 3 consists of a thorough review of the data mining methodology. Section 4 describes the results and discusses the various evaluations carried out and, finally, the paper comes to an end with a conclusion and future work in Section 5.

II. RELATED WORK

In the paper [1], the Authors are applying multiple regression methods on Big mart sales data to predict the sales of a product in a particular outlet. In this paper, they are finding the factors behind the sales of a product and store that is impacting the sales of that product. Mean Absolute Error (MAE) is used to evaluate the performance and accuracy of the model. The two-level statistical model performed well in comparison to the rest of the single model predictive techniques. This stacking technique applied to machine learning algorithms was not much useful while applying to the dataset used in this paper as the performance was decreasing in comparison to the single-level technique.

In the second research paper [2], SVR and artificial neural networks are used to predict the price of the car. On applying just one ML algorithm, the accuracy was 50%. But, on applying two algorithms, it was increased to 98.23%. This approach is giving good results for car predictions, provided all the factors used while predicting are present. If a few less important factors are missing, the results are degrading.

In the paper [3], the authors come up with a method of Multi-scale affinity propagation(MSAP) in which they were clustering houses (dividing the city into different price zone) based on their landmark and the facility. Then in each aggregation of houses, Linear regression is used to predict the reasonable price. MSAP has shown a remarkable effect on a prediction of a reasonable price. There are more factors other than the distance from the landmarks which can be the deciding factor while price prediction of the rental property.

In the research paper [4], the authors are predicting the price of the rental property using various machine learning models like linear regression, tree-based, SVR, KNN, and neural networks and comparing all these to get the best prediction model. MSE, MAE, and R2 scores as the evaluation methods used. SVR and neural networks outperformed amongst all the machine learning techniques with SVR giving an R2 score of 69% and an MSE of 0.147 on the test dataset.

In paper [5], the price prediction model for second-hand cars was built using multiple linear regression, random forest regression, gradient boosting regression. Amongst all gradient boosting outperformed with an MAE of 0.28. It is well describing the objective of this paper to predict the car price using random forest regression. KNN regression could also have been applied for more effective results as shown in the undergoing paper for the car price prediction model.

In paper [6], the model is created to predict the price of the rental property using three models namely support vector machines (SVM), Random forest, and gradient boosting machine. Random forest and gradient boosting are performing well in comparison to the SVM model. But using a more conventional model like linear regression in predicting the price of rental property is giving comparatively good results. And it can be seen in the ongoing research paper.

In paper [7], a general comparison is made between the various machine learning algorithms like Artificial neural networks, Decision trees, KNN algorithm, Logistic regression, Random Forest, and support vector machines. The authors of this paper have divided the areas according to the machine learning models where it could be applied and get good results. As per the research paper, a vast domain is missing where these models could be applied.

In Paper [8], the performance of multiple machine learning algorithms was compared for different datasets. Authors found out that using predictive models as a benchmark is not perfect unlike the paper [7]. The conclusion of this paper fits the ongoing research paper as in applying the conventional methodologies like linear regression on such complex data like Airbnb are outperforming the non-conventional ones.

In paper [9], the stock price is predicted using the ensemble classifier called GASVM based on support vector machines with a generic algorithm (GA). The paper shows that GASVM outperformed all the other classic machine learning algorithms as it was having a prediction accuracy of 93.7%. This technique of predicting is quite fascinating and can be used while dealing with economical data. This paper was referenced more to see the usage of the algorithm than to apply this technique in the ongoing project as it is beyond scope of this project.

In the research paper [10], Linear regression and the random forest is used to create the car price prediction, model. Authors have applied feature selection methodologies to get a hierarchy of the features while deciding the car price. From this paper, feature selection of the cars is used and then to apply those features in the models. This approach will help optimize models from the first place.

In the paper [11], the linear regression model is used to predict the future sales of a big superstore. This model is having an accuracy rate of 84% approximately. The approach of applying the linear model is not used to forecast the sales of liquor in the ongoing paper as it is not in the scope of applying the same model twice on the datasets., instead a support vector regressor is used which is underperforming comparatively.

In the paper [12], machine learning models are used to predict the price of the Airbnb listings in New York City. XGBoost outperformed all the models with an accuracy of 62% approximately. The technique from this paper is applied in the ongoing project and the results were a bit different as the linear model was performing better than XGBoost for this dataset.

In the paper [13], models for data mining are explained sequentially. Authors have compared KDD, CRISP-DM, and SEMMA and have given a brief understanding of them.KDD is the complete and accurate model to be undertaken for the data mining process and is followed by most researchers in this field. As per the project, KDD is the most suitable methodology to apply to the ongoing project.

III. DATA MINING METHODOLOGY

Data mining plays an important role in the knowledge discovery process. KDD (Knowledge Discovery from Databases) is used in this project for the data mining process. It consists of five stages with the ability to go back to the previous stage if necessary, which is perfect for this project. KDD refers to the exploration of data knowledge and emphasizes the high level of the particular data mining process [13]. It is used to obtain hidden knowledge and needs that the goals of the project and the underlying business criteria have already been established. Figure 1 shows the KDD methodology.

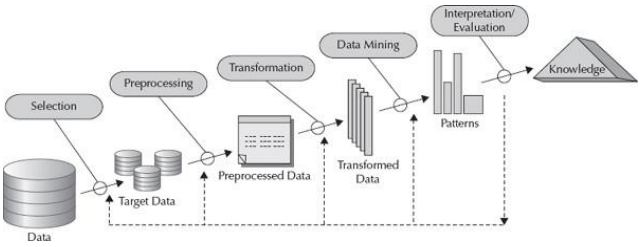


Figure 1: KDD Method

Below are the steps of the KDD process and its application in the undertaken project.

A. Selection of the Datasets

This is the first stage of KDD which consists of the development of a target data set from the available data sources. During this process, the emphasis is on the collection of attribute subsets and the sampling of data to minimize the number of records to be used during the remaining steps.

1) Iowa Liquor Dataset:

The Iowa Department of Commerce allows any store that sells alcohol in bottled form for off-premises have to possess a

Class "E" liquor license. All alcoholic purchases made by stores registered with the Iowa Department of Commerce shall be logged in the Commerce Department System, which is then be released by the Iowa Department of Commerce as open data. This dataset contains 18 columns and 270955 rows in total which was 1 million rows initially. Rows were selected based on the sales year in which the transaction was done. “Sale (Dollars)” is the target variable. This target variable would be predicted based on the remaining 17 predictors. Figure 2 displays the structure of the Iowa Liquor dataset.

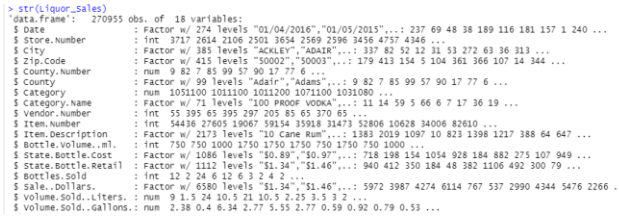


Figure 2: Structure of Liquor Dataset

2) *Used Cars Dataset:*

Craigslist is the largest collection of used cars for sale in the world. This dataset contains the car used on Craigslist throughout the United States. This dataset contains 458260 rows and 25 columns, out of which the “Price” column is the target variable. Figure 3 displays the structure of the used cars dataset. Some important predictors are as follows:

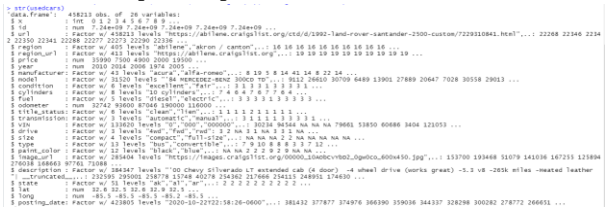


Figure 3: Structure of Used Cars Dataset

3) *Airbnb Dataset:*

Airbnb is a hospitality industry that provides travelers a more exclusive and customized place to stay and visit. Today, travelers are more attracted to Airbnb as their primary means of accommodation. This dataset contains 29 columns and 74111 columns. “Log_Price” is the target variable, which would be predicted using the available predictors. Figure 4 displays the structure of the Airbnb dataset.

Figure 4: Structure of Airbnb Dataset

The above three target datasets are now sent to the Pre-Processing phase of KDD.

B. *Data Pre-Processing*

This stage consists of cleaning and pre-processing the data set to be used for data mining. Some of the activities involved at this stage include identifying and removing noise, identifying and deciding how to handle missing data, cleaning up data anomalies, and so on. This step is very important as the unprocessed data can lead to bad models and will have incorrect accuracy.

1) *Iowa Liquor Dataset:*

The following preprocessing was done in this dataset to deal with the data anomalies in this dataset:

a. Missing Data:

Missing data in the dataset was analyzed to make the data consistent to perform future operations with it.

Figure 5: Missing data in Liquor data

As per figure 5, there are not many missing data in this dataset. “County.Number”, “County”, “Category”, and “Category.Name” has very few missing data when comparing it with the dimension of the data.

Since these missing values are few, are not very significant predictors except the "Category.Name" and cannot logically be filled with any handling methods, thereby eliminating such missing values from the data.

b. Outliers Detection:

Removing outliers from data is an important step as it can affect the prediction model negatively and can discrepant the model accuracy.

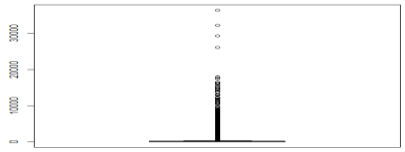


Fig 6: Showing Outliers in Target Variable

Figure 6 displays the outliers in the target variable. The outlier is checked in the target variable. The target variable in this dataset is continuous and the data type of it should be numeric rather than being the factor type which it is initially with ‘$’ sign concatenated to it. Transforming the data type of it to numeric and removing the ‘$’ sign from it. After the data type is changed, outliers from it are removed. Figure 7 displays the boxplot after removing the outliers.

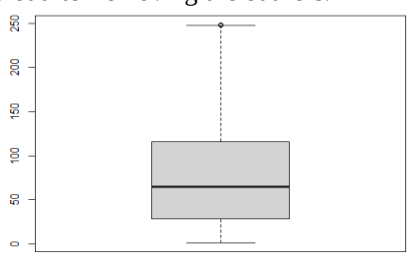


Fig 7: After removing Outliers in Target Variable

This dataset is pre-processed and is ready for a few transformations in it.

2) *Used Cars Dataset:*

Below are the pre-processing performed in this dataset.

a. Missing Data:

Missing data in the used cars dataset was analyzed to make the data consistent and ready for predicting the second-hand car price.

```
> sapply(usedcars,function(x) sum(is.na(x)))
          X          id         url      region  region_url       price        year manufacturer       model   condition   cylinders
          0           0           0           0           0           0        1050       18220        4846      192940      171140
       fuel    odometer title_status transmission         VIN       drive        size        type  paint_color   image_url description
       3237       55303        2577        2442      187549      134188      321348      112738      140843          28          70
      state         lat        long posting_date
          0        7448        7448          28
```

Figure 8: Missing data in Used-car data

We dealt with the missing values for this dataset in a bit different way by applying the domain knowledge of this dataset. As per figure 8, removing rows with Columns having few missing data such as year, model, fuel, transmission, drive, type, and paint_color. It was better to remove these rows as filling columns with imputed values was meaningless and it would lead to the bad price prediction model.

Now, the predictors such as condition and cylinders were having a fair amount of missing data, in which the number of cylinders was filled by imputing the values through linear regression, but the strategy was not effective and the model accuracy was very low, hence dropped rows with null values. Now, the predictor "Condition" was tried imputed by Multiple Logistic regression, but the accuracy of it was low. Hence, removed the rows with null values in cylinders.

Now, dealing with the predictor variable "size", as it was having too much of missing values and it cannot be imputed with predicted values, removing the entire column as it is not an important predictor.

Checking for the zero price of the car - As it is bad data and will not contribute in building the model for this dataset. Hence, removing it from the data frame.

```
> nrow(usedcars[usedcars$price ==0,])
[1] 33753
> usedcars <- usedcars[usedcars$price !=0,]
```

Figure 9: Number of rows having zero price

b. Outliers Detection:

Removing outliers from the target variable to better create the model.

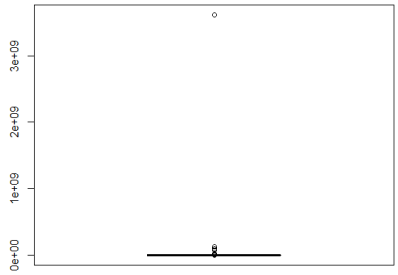


Figure 10: Showing Outliers in Target Variable

Figure 10 display the outliers in the target variable and needs to be removed to have the correct model. Hence removing outliers from the target variable.

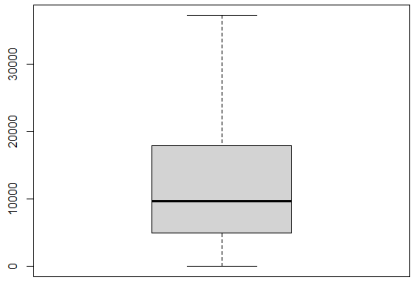


Fig 11: After removing Outliers from Target Variable

Data will now be sent to the transformation phase.

3) *Airbnb Dataset:*

Airbnb data is pre-processed to make it ready for the next stages and to build the price prediction model for the rental property.

a. Missing Data:

Missing data in any dataset is discovered to remove the inconsistencies from it.

Since the Airbnb dataset is too large, filtering the data on New York City and then discovering the missing values in the dataset.

```
> sapply(AirbnbData,function(x) sum(is.na(x)))
                      id                log_price            property_type                room_type                amenities             accommodates
                       0                        0                        0                        0                        0                        0
               bathrooms                 bed_type      cancellation_policy           cleaning_fee                     city              description
                      99                        0                        0                        0                        0                        0
            first_review host_has_profile_pic host_identity_verified host_response_rate              host_since         instant_bookable
                    6858                      176                      176                     9960                      176                        0
             last_review                 latitude                longitude                     name             neighbourhood        number_of_reviews
                    6832                        0                        0                        8                        0
review_scores_rating              thumbnail_url                  zipcode                 bedrooms                     beds
                    7321                     2415                      446                       47                       65
```

Figure 12 Missing Data in Airbnb

Looking at the null values in New York City Airbnb data, it is visible that very little data is missing and the missing values are present in less important predictors, which will be removed in the data transformation phase later on.

"review_score_rating" is an important predictor, which has a good amount of missing values. Rows with missing values are removed from the data.

The price of the properties which are 0 is removed from the dataset as it is bad data and will lead to erroneous model creation.

b. Outliers Detection:

Outliers detection is done as it is done in the previous datasets and is removed from the target variable.

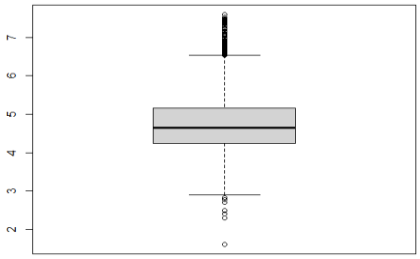


Figure 13: Outliers detection in the target variable

Hence, removed those outliers and plotted the boxplot again to verify it.

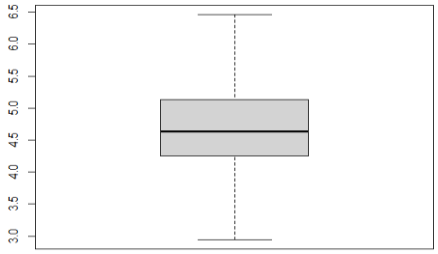


Figure 14: After removing Outliers from the target variable

Now, Data is ready for the transformation phase of KDD methodology.

C. Data Transformation

After pre-processing of all the three datasets used in this project, now they will go under the transformation phase of KDD, which is the third phase of this methodology. This stage consists of defining and applying the appropriate data transformations. This may include the use of different binning and aggregation approaches to be used in data transformation and data dimensionality reduction.

1) Iowa Liquor Dataset:

Now, as the data is pre-processed, the transformation of data is done to make the data ready for applying sales prediction models. Various transformations done on the Liquor dataset are as follows:

Date Predictor:

On loading the dataset, initially, the date column was in factor type format, it was converted into the Date format of R.

State.Bottle.Cost and State.Bottle.Retail Predictors:

Initially, both the predictors were of factor format with "$" concatenated with the values. It was transformed in the numeric data type and the $ sign was removed from the values.

Outliers:

The Predictor variable (State.Bottle.Cost and State.Bottle.Retail) are removed. Now, it is ready to be given to the model to create the sales prediction model of Iowa liquor.

2) Used Cars Dataset:

Transformation of used cars dataset is done to give it to the model to create a car price prediction model. It undergoes various transformation phases to fit it into the applied machine learning model. Below are the transformations done on this dataset.

Outliers from the main predictor variable "Odometer"

Odometer outliers were removed from the dataset.

Checking the mean price across each state of the United States

Aggregation was performed on the "price" of used cars and the mean price of each state of the United States was calculated to see if the price is dependent on the location predictors.

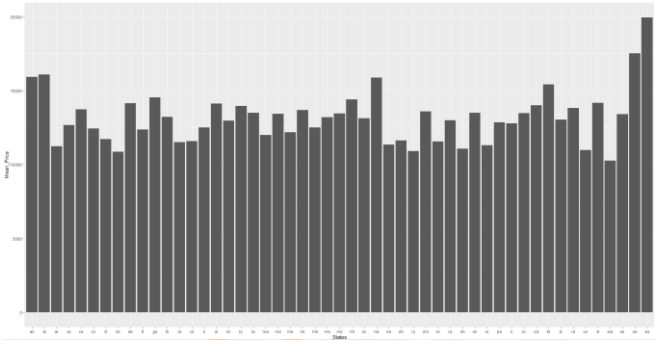


Figure 15: Mean price across each state of the united states

According to figure 15, the mean price across each state is nearly the same and the price does not differ according to the location. Hence, removing all the columns regarding the locations from the dataset.

Columns that were removed from the analysis are region, region_url, state, long, lat.

Few more unimportant columns were removed from the dataset as they were not contributing anything to create the price prediction model. Columns such as id, url, VIN, posting_date, image_url, description, and title_status.

The "manufacturer" column is also removed from the analysis as we have a "model" predictor in the dataset and it represents the same thing. Hence, removing the "manufacturer" column from the dataset.

Checking used-car "price" against "odometer"

Plotting scatter plot to see the distribution of price against the odometer reading. There is a strong relationship between these two variables as per the domain knowledge. Odometer reading increases, price decreases.

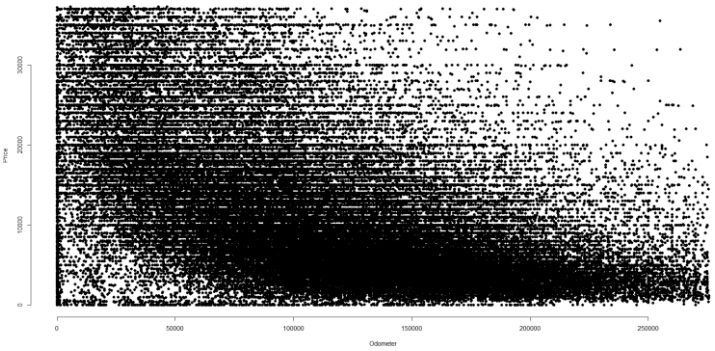


Figure 16: Scatter plot of Price against Odometer Reading

Figure 16 reveals that a few new vehicles were sold free of charge. Hence, these are bad data and needs to be removed from the dataset as it can negatively affect the model. To remove those bad data, Setting a threshold value of 5000

(Price + Odometer) and filtering the data that are violating the filter condition.

Checking for very Older cars

Very Old cars in the dataset can affect the model creation as it can have a relatively higher price in the name of vintage cars as shown in figure 17. Hence, removing rows with a year less than 1960.

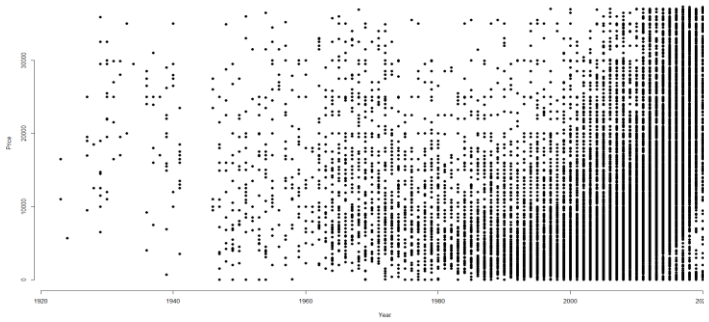


Figure 17: Scatter plot against Year and Price

All the generalized transformations are done on the dataset. A machine learning model requires some specific transformations, it will be explained in the Model Creation part of the report.

3) Airbnb Dataset:

Airbnb dataset is transformed and made ready for the machine learning models. To achieve this, various transformations are performed on the dataset.

Removing Bad data:

Few domain-specific transformations need to be done on the dataset. The number of "beds", "accommodates", "bathrooms" and, "bedrooms" cannot be zero for a rental property as it does not makes any sense for the customers to rent such properties. Hence, removing those rows having value as zero for these columns.

Re-leveling the predictor "host_identity_verified" and "instant_bookable":

Initially, the column was having values as "f" and "t" which was transformed into "False" and "True" and was leveled "True" and "False".

Extracting year part from "host_since" predictor:

The year part of this column is stored in a new column named "year_host".

Calculating served days as the host on the Airbnb platform:

It is done by extracting the year part from "last_review" and storing it in a column named "year_end". Likewise, from the "first_review" predictor and storing it in the "year_start" column. Now, taking the difference of both these columns and storing it in "diff_dates" columns as the number of days served on the Airbnb platform.

Removing a few unwanted and temporary created columns:

Removed a couple of variables like id, thumbnail_url, host_has_profile_pic, description, first_review, last_review, city (as filtered on New York City), zipcode, host_response_rate, and a couple of temporary variables.

Merging factors of bed_type column:

Initially, this column was having five factors having values of bed_type in the property. It was transformed into two factors named "Real Bed" and "Other".

Removing levels with few rows in "cancellation_policy":

Removed levels "super_strict_30" and "super_strict_60" levels from the cancellation_policy variable as it was having very few rows.

Merged 35 levels of "property_type" variable into 6 levels:

There were many levels initially which resembles the same property type. Hence, all those were merged into Six levels in the property_type variable named Apartment, Villa, Hostel, B&B, Timeshare, and Other.

Calculate the average price of the neighborhood:

The average price of the room of the neighborhood is calculated in which the property is situated. It is achieved by calculating the price per room (log_price/bedrooms) and was stored in the variable price_per_room. Afterward, the average price per room of the neighborhood was calculated by taking the mean value of price_per_room on grouping the "neighborhood" column and storing it in a new dataframe. Later on, the newly created dataframe is merged with the original on "neighborhood" as the primary key.

Now, the transformed data is ready to apply in the machine learning model.

D. Data Mining

Data mining is the fourth phase of the KDD methodology. This stage consists of choosing the necessary data mining algorithms for the chosen data collection. This data mining algorithm looks for patterns that might appear in the data. Relevant parameter settings would need to be determined to ensure optimum algorithm operation.

1) Iowa Liquor Dataset:

The goal of using this dataset is predicting future sales on present sales value. Since our target variable i.e. "Sale..Dollars." is a continuous variable, applying the regression data mining method.

From various regression models, Support Vector Regression (SVR) is applied to this dataset to create the sales prediction model. It is a supervised machine learning model and works the same as SVM (Support Vector Machines) with few

differences. It employs nonlinear mapping to construct a linear model in the feature space [6].

$Y_{i=} W.K(x_i,x)+b$

Now, the SVM model is implemented using the "e1071" package in R. Sampling of the dataset was done. Data was split into train and test data. Training data was supplied along with the target and predictor variable into the SVM() function. Using the model created, sales value was predicted for the test data. Plotted the predicted value against the actual value of test data. After that, the Value of W and b is calculated. RMSE of the model is seen and if the RMSE value is very high, the SVR model is tuned. It will be explained in a more detailed manner in the evaluation section of the model.

2) Used Cars Dataset:

The goal of this data set was to predict the used car price. Our target variable is a continuous variable, hence applying regression data mining methods.

On this dataset, KNN for regression and Random Forest is applied.

KNN machine learning Model:

To apply the KNN algorithm for regression on a dataset, certain model-specific transformations need to be done to get the correct model. Following are the processes carried out to apply the KNN on this dataset:

- Predictors and target variables were separated and stored in two different variables.
- Scaling of predictor "odometer" was done to scale the data of this variable.
- Dummy coding of predictors which are factors of more than two levels such as "condition", "cylinders", "fuel", "transmission", "drive", "type", "paint_color" is done and are stored in different new variables.
- Since cylinder, fuel, transmission, and type predictors were having "Other" as one of the levels. Hence, renamed this level to avoid duplication and error later on.
- Newly created variables having dummy code of factors are bound with the dataframe having predictor variables created earlier in the first step.
- Random sampling is done, data is divided into train and test data.
- Train data with its outcome variable is passed with test data on which prediction is to be done into the knn.reg() function.
- The value of k is adjusted and accuracy is calculated.

Random Forset machine learning model:

To apply the Random Forest model on this dataset, few transformations are done as per the model before passing the data in the model. randomForest package was installed and used to implement this machine learning model. Below are the steps carried out to implement this model.

- Factor predictors such as "condition", "cylinders", "fuel", "transmission", "drive", "type" and, "paint_color" are converted in the numeric data type.
- Predictors and target variables were separated and stored in two different variables.
- Random sampling is done, data is divided into train and test data.
- Train data with its outcome variable is passed with the number of maxnodes and ntree into the randomForest() function.
- Value of maxnodes and ntree are adjusted and the most accurate model is selected.

3) Airbnb Dataset:

The aim of using this dataset is to predict the price of the rental property per night. The target variable "log_price" is a continuous variable and it is the rent of the property per night. Being a regression variable, We have applied a Linear Regression analysis and XGBoost regression model to predict the log_price.

Multiple Linear Regression:

To apply Linear regression, there is no such special transformation required. Following are the processes carried out to implement this model.

- Random sampling is done, data is divided into train and test data.
- The formula of the regression model is created.
- Train data along with the formula is passed into the lm() function.
- Prediction based on the created model is done on the test data.
- Predicted and actual values of the price are passed into the GainCurve plot to see the accuracy according to the Gini score.
- Cross-validation is done to get the number of trees

XGBoost Regression:

Applied XGBoost regression on the Airbnb dataset to predict the log_price.

Following are the steps that are performed to implement this model.

- Random sampling is done, data is divided into train and test data.
- Cross-validation is done to calculate the number of trees in train and test data. It gives the index of minimum RMSE as a number of trees in the train and test data.
- Now, with the number of trees, training data parameters like depth of the tree, eta is passed in the xgboost() function to create a model.
- Price of the Airbnb is predicted according to the model created.
- The accuracy of the model is tested if it is acceptable or needs to be analyzed.

Evaluation of all these applied models in the above three datasets will be done in the Evaluation section in detail along with the impact of parameterization on the results. Also, the sampling methods used while modeling will be explained.

IV. EVALUATION

This stage consists of evaluating the results and outputs of data mining algorithms applied in the above section to see if acceptable and usable trends have been found. Various visualizations and predictive assessments will be produced as part of the evaluation phase. Below are the evaluations of each model created in the data mining phase of KDD.

1) Iowa Liquor Dataset:

In this dataset, SVR was applied to create a model that could predict future sales based on the present sales. Below is the evaluation of the model created.

SVR: Calculating RMSE

In this Model, training and test data were passed by using the Random Sampling Technique. Each entity is included solely by chance and each member of the population is having an equal chance of being included in the sample.

Initially, the RMSE of the created model was very high and was equal to 82% approximately, which indicates that predicted values were very far from the actual values.

Tuning of SVR (Support Vector Regression)

It was done to increase accuracy and sensitivity analysis was performed to have a better model. In this analysis, a lot of models is trained with various allowable error and cost parameter shown in figure 18.

```
> OptModelsvm=tune(svm, Sale..Dollars.~.-Date-Store.Number-Category.Name-City-Zip.Code-County.Number-County-Vendor.Number-Item.Number-Item.Description, da
ta=Liquor_train,ranges=list(elsilon=seq(0,1,0.1), cost=1:100))
> print(OptModelsvm)

Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 elsilon cost
       0    4

- best performance: 11.95359
```

Figure 18: Tuning the SVR Model

The above R code shown in figure 18, tunes the SVR model by evaluating the performance of 1100 models i.e. for every combination of allowable error and cost parameter. Value of epsilon and cost are 0 and 4 respectively in the optimized model.

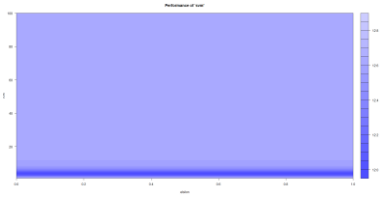


Figure 19: Plot of the Optimised SVM model

From the collection of optimized models, selecting the best model. Figure 20 displays the selection of the best model.

```
> BstModel

Call:
best.tune(method = svm, train.x = Sale..Dollars. ~ . - Date - Store.Number - Category.Name - City - Zip.Code - County.Number -
    County - Vendor.Number - Item.Number - Item.Description, data = Liquor_train, ranges = list(elsilon = seq(0, 1, 0.1),
    cost = 1:100))


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  4
      gamma:  0.1428571
    epsilon:  0.1

Number of Support Vectors:  168
```

Figure 20: Selecting Best Model

The Best Model of SVM is having an RMSE of 3 % approximately as shown in Figure 21, which is using the "radial" kernel of SVR. RMSE has been improved a lot by tuning the SVM and using this kernel in comparison to the previous model having different kernels of SVR.

```
> RMSEBst=rmse(PredYBst,Liquor_test$Sale..Dollars.)
> RMSEBst
[1] 3.370126
```

Figure 21: RMSE of Best Model

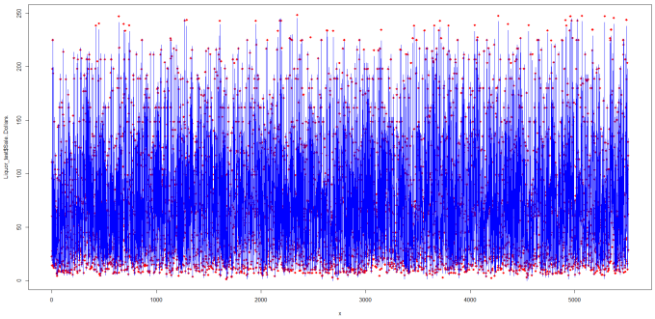


Figure 22: Actual(Red) vs Predicted(Blue)

2) Used Cars Dataset:

Initially, this dataset was cleaned and transformed rigorously to create an optimized model to predict the price of used cars. Below are the evaluations of the model applied.

KNN Model

In this model, the random sampling method was used to create a Train and Test dataset. This technique was used to create an equal amount of chance for each data in the dataset to be included in the sample.

Initially, we have to decide the number of k. There are many rules to determine the number of k, but the most considered one is taking the square root of the number of observations which is near about 300 for this case. But considering this value of k, the model accuracy was 79 % approximately and as the K value is decreased, the accuracy of the model is increasing. But, a small value of K means that noise would be influencing the result and a large value is computationally expensive.

But in this case, as there is not a huge difference and a smaller K may utilize more subtle patterns, choosing the value of K = 17 which is having an accuracy of 84% approximately and is not computationally expensive.

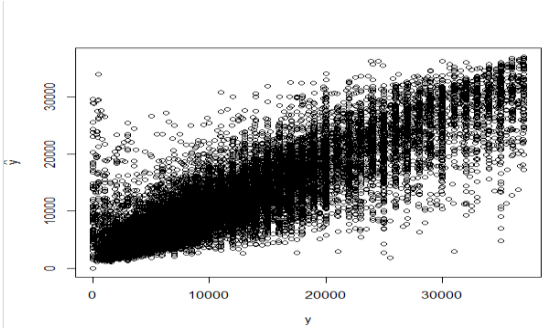


Figure 23: Actual vs Predicted

The graph shows, the values are very well predicted as these are on the y=x line.

Random Forest

In this model, training and test data are created using the stratified sampling method to divide the dataset into mutually exclusive and exhaustive subgroups. It ensures the diversity of the sample created with similar variance.

Initially, the created model was having an R^2 value of 74% approximately with parameter ntree as 100 and maxnode=20 as shown in figure 24. It is an acceptable accuracy percentage.

```
> print(paste0('R2: ' ,caret::postResample(predictions , y_test)['Rsquared'] ))
[1] "R2: 0.749035043057338"
> print(paste0('RMSE: ' ,caret::postResample(predictions , y_test)['RMSE'] ))
[1] "RMSE: 5064.80771672949"
```

Figure 24: Accuracy of Random Forest Model

But, still trying to tune the model to get a more significant model. Plotting the initial model with error rate across trees as shown in figure 25. As the number of trees are increasing, the error rate is decreasing.

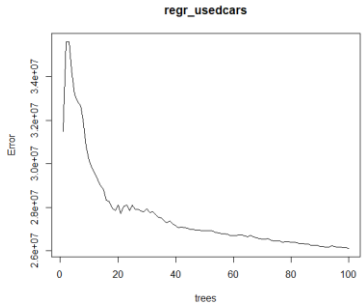


Figure 25: Error rate of Model

Finding the number of trees with the least error, and re-running the model by changing the number of trees to get a more accurate prediction. The number of trees should not be extremely more or less, it should be in between the extremes somewhere because it can lead to overfitting of data. Selected the number near to the accuracy of the model is increased as the RMSE is decreased by a small percentage with the same R^2 value.

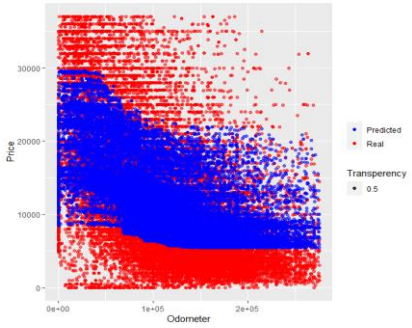


Figure 26: Actual vs Predicted

Comparing both the models Random Forest and KNN applied on this used cars dataset, KNN outperformed Random Forest with the accuracy of 84% approximately.

3) Airbnb Dataset:

This dataset is used to create a price prediction model using Multiple Linear regression and the XGBoost model.

Multiple Linear Regression

It is the simplest regression and effective technique to do regression analysis. In this regression, a random sampling technique is used. It is used to randomly sample the dataset with each row having the same probability of making it into the sample data.

Using the broom library in R, computed the value of R^2 which is 69.6%. It shows that property type and average neighborhood price affects most the price of the rental property.

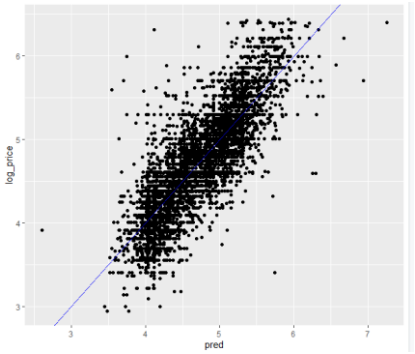


Figure 27: Actual VS predicted Linear Regression

Evaluating the accuracy rate of Linear Regression other than R^2. Calculating the Gini Score of the curve.

Figure 28: GainCurve Plot

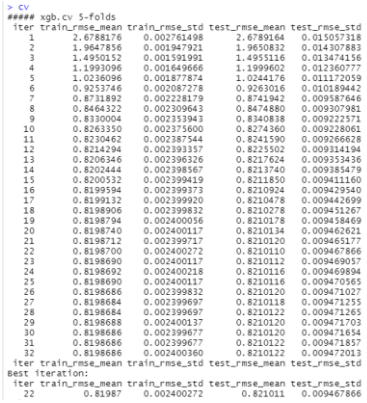
According to figure 28, the Gini score is close to 1, which shows that predictions sort in the exactly same order. Gini coefficient close to 0 represents that model is sorting poorly.

XGBoost Regression

The random sampling technique is used to create train and test data to pass to the model to create a prediction model.

To create a model through XGBoost, a cross-validation technique is used to search for the index of the number of trees at which the RMSE is minimum, and accordingly, it is passed into the model to create a model.

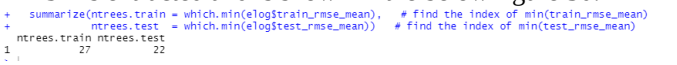
Cross-validation is done and figure 29 displays the evaluation log and the Best iteration among that.



```
> cv
#### xgb.cv 5-folds
 iter train_rmse_mean train_rmse_std test_rmse_mean test_rmse_std
    1       2.6788176    0.002761498       2.6789164   0.015057318
    2       1.9647856    0.001947921       1.9650832   0.014307883
    3       1.4950152    0.001591991       1.4955116   0.013474156
    4       1.1993096    0.001649666       1.1999602   0.012360777
    5       1.0236096    0.001877874       1.0244176   0.011172059
    6       0.9253746    0.002087278       0.9263016   0.010189442
    7       0.8731892    0.002228179       0.8741942   0.009587646
    8       0.8464322    0.002309643       0.8474880   0.009307981
    9       0.8330004    0.002353943       0.8340838   0.009222571
   10       0.8263350    0.002375600       0.8274360   0.009228061
   11       0.8230462    0.002387544       0.8241590   0.009266628
   12       0.8214294    0.002393357       0.8225502   0.009314194
   13       0.8206346    0.002396326       0.8217624   0.009353436
   14       0.8202444    0.002398567       0.8213740   0.009385479
   15       0.8200532    0.002399419       0.8211850   0.009411160
   16       0.8199594    0.002399373       0.8210924   0.009429540
   17       0.8199132    0.002399920       0.8210478   0.009442699
   18       0.8198906    0.002399832       0.8210278   0.009451267
   19       0.8198794    0.002400056       0.8210178   0.009458469
   20       0.8198740    0.002400117       0.8210134   0.009462621
   21       0.8198712    0.002399717       0.8210120   0.009465177
   22       0.8198700    0.002400272       0.8210110   0.009467866
   23       0.8198690    0.002400117       0.8210112   0.009469057
   24       0.8198692    0.002400218       0.8210116   0.009469894
   25       0.8198690    0.002400117       0.8210116   0.009470565
   26       0.8198686    0.002399832       0.8210120   0.009471027
   27       0.8198684    0.002399697       0.8210118   0.009471255
   28       0.8198684    0.002399697       0.8210122   0.009471265
   29       0.8198688    0.002400137       0.8210120   0.009471703
   30       0.8198686    0.002399677       0.8210120   0.009471654
   31       0.8198686    0.002399677       0.8210122   0.009471857
   32       0.8198686    0.002400360       0.8210122   0.009472013
 iter train_rmse_mean train_rmse_std test_rmse_mean test_rmse_std
Best iteration:
 iter train_rmse_mean train_rmse_std test_rmse_mean test_rmse_std
   22         0.81987   0.002400272       0.821011   0.009467866
```

Figure 29: Cross-Validation

The number of trees in Test and Train with minimum RMSE is extracted and is shown in the below figure 30.



```
+   summarize(ntrees.train = which.min(elog$train_rmse_mean),   # find the index of min(train_rmse_mean)
+             ntrees.test  = which.min(elog$test_rmse_mean))   # find the index of min(test_rmse_mean)
  ntrees.train ntrees.test
1           27          22
```

Figure 30: Number of trees in Train and Test

Now, the XGBoost model is executed with the parameter calculated in the above figure 30 for the accurate model. This model for this dataset did not perform well and have an accuracy of just 20% approximately and have a huge RMSE of 0.8.

According to the applied models on Airbnb, Multiple Linear Regression outperformed the XGBoost model with an accuracy of 70% approximately.

V. CONCLUSION AND FUTURE WORK

This paper carried out a comparative analysis of the results of 5 machine learning models using the KDD approach of data mining. All 5 machine learning models have been successfully performed on three separate data sets and have achieved satisfactory results in creating dataset-specific models. The KNN regression model, Random Forest, and Multiple Linear Regression gave adequate results on the performed datasets. The KNN model had the highest accuracy of 84 percent and the XGBoost model had the lowest accuracy of 20 percent on the applied datasets. SVR (Support Vector Regressor) was tuned by evaluating the performance of 1100 models and the model with lower RMSE was chosen as the final model for predicting future sales of liquor. Multiple Linear regression outperformed XGBoost in predicting the price of the rental property and the KNN model outperformed the Random Forest model in predicting second-hand cars price. The evaluation of models was done using R^2, RMSE, and Gini Score.

From this analysis, future work can be built by fine-tuning each model parameter [5]. A two-level machine learning model could be applied to the Airbnb data to get more appropriate results. The next steps in sales predictions are to add deep learning architectures in the created models. To extend these created models on various other datasets to measure the accuracy. To predict the car price more effectively, PCA of car features is to be done and accordingly, the factors can be used in implementing the model of the used car dataset.

REFERENCES

[1] K. Punam, R. Pamula and P. K. Jain, "A Two-Level Statistical Model for Big Mart Sales Prediction," 2018 International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, Uttar Pradesh, India, 2018, pp. 617-620, doi: 10.1109/GUCON.2018.8675060.

[2] Gegic, Enis, et al. "Car price prediction using machine learning techniques." TEM Journal 8.1 (2019).

[3] Y. Li, Q. Pan, T. Yang, and L. Guo, "Reasonable price recommendation on Airbnb using Multi-Scale clustering," 2016 35th Chinese Control Conference (CCC), Chengdu, 2016, pp. 7038-7041, doi: 10.1109/ChiCC.2016.7554467.

[4] Rezazadeh, Pouya & Nikolenko, Liubov & Rezaei, Hoormazd. (2019). Airbnb Price Prediction Using Machine Learning and Sentiment Analysis.

[5] N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya and P. Boonpou, "Prediction of prices for used car by using regression models," 2018 5th International Conference on Business and Industrial Research (ICBIR), Bangkok, 2018, pp. 115-119, doi: 10.1109/ICBIR.2018.8391177.

[6] Winky K.O. Ho, Bo-Sin Tang & Siu Wai Wong (2020) Predicting property prices with machine learning algorithms, Journal of Property Research, DOI: 10.1080/09599916.2020.1832558

[7] Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 1310-1315.

[8] V. Tsoukas, K. Kolomvatsos, V. Chioktour and A. Kakarountas, "A Comparative Assessment of Machine Learning Algorithms for Events Detection," 2019 4th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM), Piraeus, Greece, 2019, pp. 1-4, doi: 10.1109/SEEDA-CECNSM.2019.8908366.

[9] Kumar, I. et al. "A Comparative Study of Supervised Machine Learning Algorithms for Stock Market Trend Prediction." 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (2018): 1003-1007.

[10] Madhuvanthi, K. & Kailasanathan, Nallakaruppan & N C, Senthilkumar & Somayaji, Siva. (2019). Car Sales Prediction Using Machine Learning Algorithmns. International Journal of Innovative Technology and Exploring Engineering. 8.

[11] G. T., R. Choudhary and S. Prasad, "Prediction of Sales Value in Online shopping using Linear Regression," 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-6, doi: 10.1109/CCAA.2018.8777620.

[12] A. Zhu, R. Li and Z. Xie, "Machine Learning Prediction of New York Airbnb Prices," 2020 Third International Conference on Artificial Intelligence for Industries (AI4I), Irvine, CA, USA, 2020, pp. 1-5, doi: 10.1109/AI4I49448.2020.00007.

[13] Shafique, Umair & Qaiser, Haseeb. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research. 12. 2351-8014.