MACHINE LEARNING

MODULE NAME : MACHINE LEARNING
SUBJECT CODE : ST3189
UOL STUDENT NUMBER : 210494123

Page Count : 10 (COVER PAGE, TABLE OF CONTENTS AND REFERENCES EXCLUDED)

Table Of Contents

Task : Unsupervised Learning

Introduction : Machine learning (ML) techniques, including unsupervised learning, analyzing and clustering unlabeled datasets, uncovering hidden patterns or data clusters autonomously. These algorithms aid tasks like clustering and dimensionality reduction, empowering data-driven decision-making and enhancing operational efficiency for businesses (IBM, 1995)

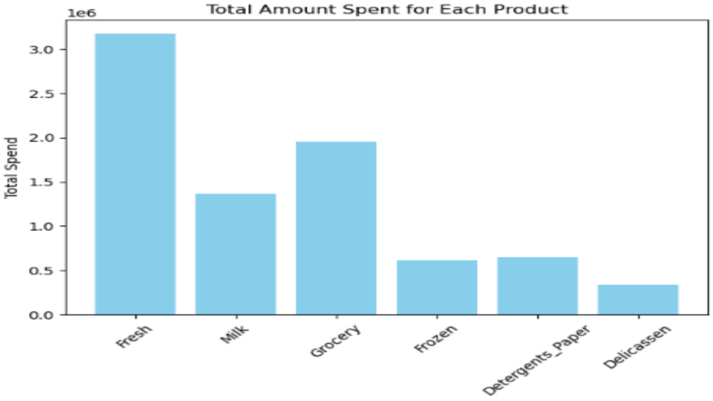
Venturing into the dynamic realm of the "Wholesale Customer" dataset from UCI Machine Learning Repository website, we embarked on a captivating exploration of customer buying habits. Armed with the analytical prowess of Principal Component Analysis and a fusion of hierarchical and non-hierarchical clustering techniques, we sought to unearth the hidden treasures of profitability within product lines, delivery channels and regional markets in the country of Portugal.

Research Questions

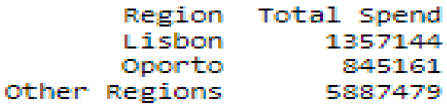
1. What product generates the most amount of revenue for the wholesale business? Assess the profitability of the product for the wholesale business.
2. Which region and channel are the most profitable in terms of sales for the wholesale business?
3. Identify the most profitable delivery channel for each region of business for the wholesale business.

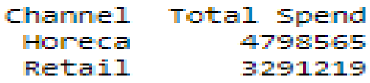
Exploratory Data Analysis (EDA)

Exploratory data analysis was carried out on the dataset in order to suit research needs and address the research questions stated above .

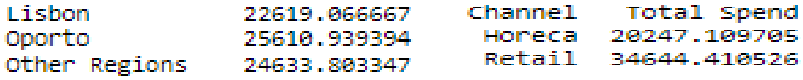


In order to identify the product generating the most revenue for the wholesale business we get the total amount spent by customers for products , we see Fresh products generates most revenue for the wholesale business . However, if we were to assess profitability of products simply using revenue generated for each product may be a subjective measure therefore, we can get the average revenue generated per transaction for each product , we see Fresh products generates 9547.4 for a transaction on average and can be inferred as the most profitable product for the wholesale business.


```
       Region  Total Spend
       Lisbon      1357144
       Oporto       845161
Other Regions      5887479
```

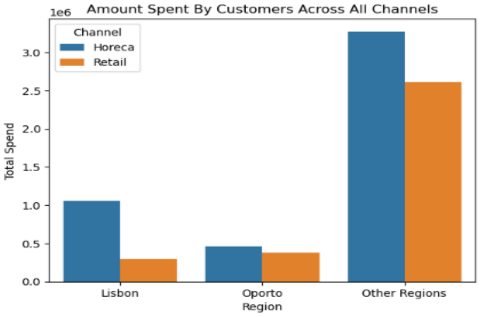

```
Channel  Total Spend
 Horeca      4798565
 Retail      3291219
```

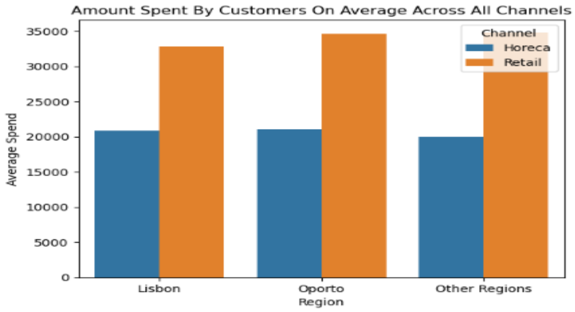
We also explore the dataset in order to identify the most profitable each region and channel for the wholesale business , if we were to look at profitability in terms of revenue generated Horeca is the most profitable delivery channel and Other regions is considered the most profitable region (figure above) . However simply using revenue to assess profitability can be subjective , therefore we use the average revenue for a transaction through each delivery channel and region to assess profitability,



```
Lisbon            22619.066667    Channel    Total Spend
Oporto            25610.939394     Horeca   20247.109705
Other Regions     24633.803347     Retail   34644.410526
```

We can now see the retail delivery channel generates more revenue on average for a transaction and Oporto region generates more revenue on average for the wholesale business.

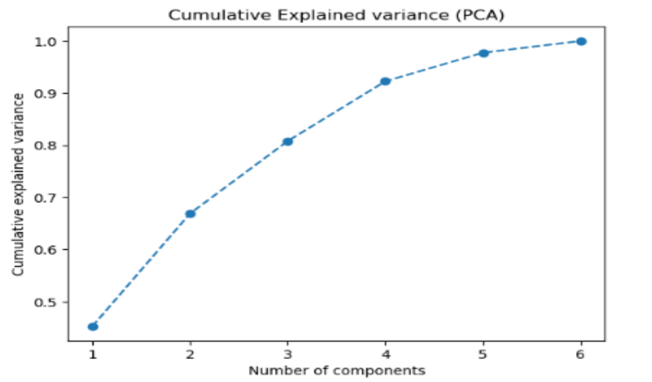




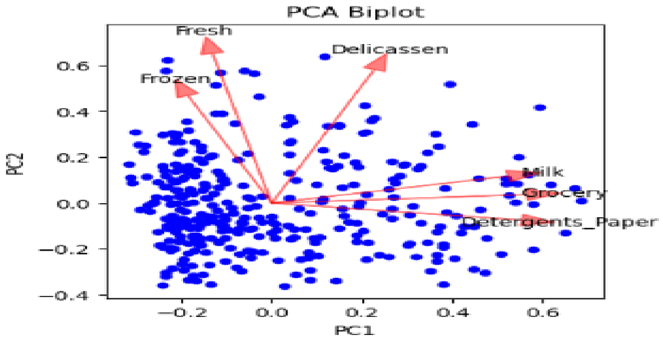
Furthermore, as per research requirement we also identify the most profitable channel for the wholesale for each business region, if we were to assess profitability of the delivery channels in terms of revenue, we see that horeca is the most dominant channel across all regions(figure above-left) , however as mentioned earlier revenue may be a subjective measure to assess profitability as some regions may have more transactions than others due to demographic reasons etc. Therefore, using the average revenue generated for each channel in each region seems a more appropriate measure for assessing the most profitable channel in each region. We see on average revenue generated retail channel shows promising signs of profitability in comparison to horeca in each region as shown in the figure above (right)

Principal Component Analysis

Principal component analysis (PCA) is a dimensionality reduction and machine learning method used to simplify a large data set into a smaller set while still maintaining significant patterns and trends. (Builtin, 2022) First, we scale the data to a normal distribution in order to ensure that the analysis is not biased by differences in feature scales, leading to more accurate and interpretable results. Principal components were constructed for metric variables , therefore 6 principal components were constructed. A cumulative explained variance graph was plotted to find the optimum number of principal components,

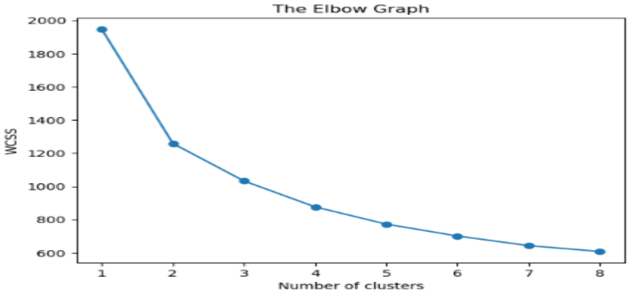


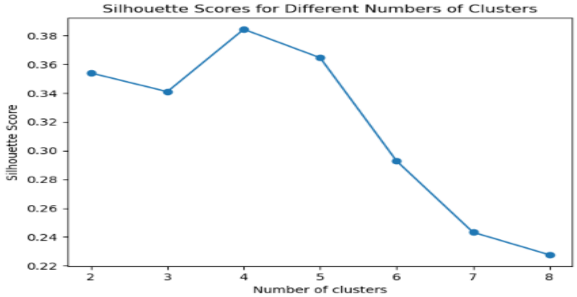
According to the graph we see that the cumulative explained variance for each component reduces as the number of components increases. The first 5 components explain more than 90% of the total variation in the dataset , therefore we will use 5 principal components for our analysis.



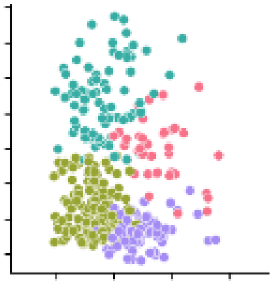
Principal component 1 and principal component 2 are used in the biplot as they capture the most amount of variation in the data. Close proximity among features suggests positive correlation between them and in the biplot above we milk, grocery and detergent paper are positively correlated. The arrows in opposite directions suggest negative correlation among features and in the case above we can see frozen goods and grocery goods are negatively correlated etc. The longer arrows represent features(milk, grocery, and detergent paper) that contribute more to the principal components and have greater influence on the structure of the data .

K-Means Clustering (Non-Hierarchical Clustering Approach)

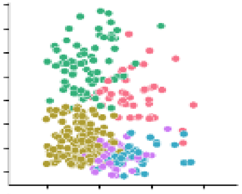




In k-means clustering, an elbow graph is a plot of the within-cluster sum of squares (WCSS) against the number of clusters. It helps to determine the optimal number of clusters for a given dataset. The elbow point is the point where the rate of decrease in WCSS slows down significantly, this point represents the optimal number of clusters for the dataset. The elbow point chosen for our analysis would be at 4 clusters as it has the highest silhouette score as shown on the diagram above(right), this would be the optimal number of clusters for k means analysis. Adding more clusters beyond this the elbow point leads to only marginal improvements in clustering performance.

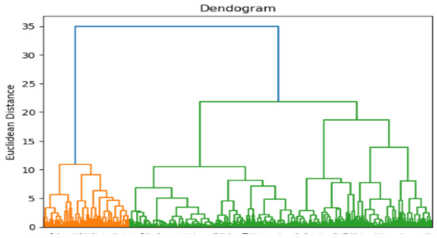


The scatterplot for 4 clusters for principal component 1 & 2 under K means clustering shows distinction with some level of interaction between data points , however we have to note that we are viewing the scatter plot from a 2D perspective. An average silhouette score of 0.38 approximately was obtained for 4 clusters which was the highest of scores compared to other cluster numbers , A higher silhouette score indicates that the object is well-matched to its own cluster and poorly matched to neighboring clusters.

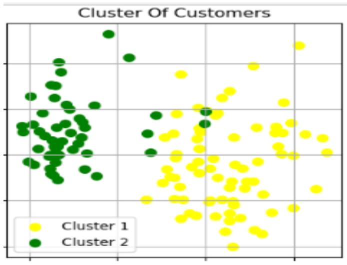


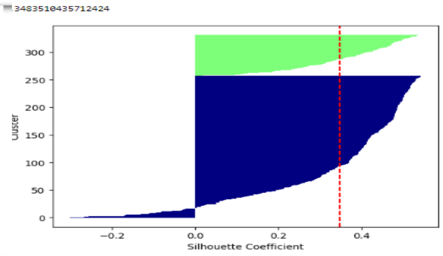
The scatterplot for 5 clusters for principal component 1 & 2 under K means clustering shows distinction with higher level of interaction between data points when compared to 4 clusters, however we have to note that we are viewing the scatter plot from a 2D perspective. An average silhouette score of 0.37 approximately was obtained for 5 clusters which was the 2nd highest of scores compared to other cluster numbers .

Agglomerative Clustering (Hierarchical Approach To Clustering)

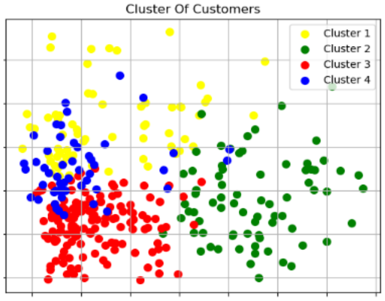


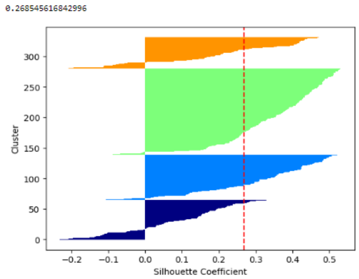
A dendrogram provides valuable information for understanding the structure of the data and deciding on the appropriate number of clusters. The dendrogram's shorter vertical lines show that the clusters or data points were merged at comparatively short distances, implying a high degree of similarity. Greater dissimilarity or lower similarity is suggested by longer vertical lines, which show that the clusters or data points were merged at larger distances. According to the dendrogram 2 clusters appear to be the optimal number of clusters , clusters being merged are more distinct from each other. This can be seen with reference to the height of the vertical lines which suggests less similarity between data points.





Under a hierarchical approach to clustering, we get the highest average silhouette score of 0.34 approximately for 2 clusters, this is in contrast to the silhouette score of 0.38 approximately for 4 clusters under a k means approach.





Under a hierarchical approach to clustering , we get an average silhouette score of 0.27 approximately for 4 clusters which is lower in comparison to the score under a k -means approach . As seen in the silhouette graph there are clusters of bars with varying heights, it may indicate mixed or overlapping clusters.

Conclusion : We see under a hierarchical approach with contrast to a non-hierarchical we appear to be needing a less number of clusters (2<4) but the difference isn't significant , however the average silhouette score calculated under a hierarchical approach appears to be slightly less than calculated under a non-hierarchical approach . Therefore, an inference can be made that 4 homogeneous clusters would be the optimal number of clusters for identifying 4 homogeneous groups for the wholesale business under a k-means clustering approach.

Task : Supervised Learning – Regression Analysis

Introduction : In supervised learning, models are trained using labelled dataset, where the model learns about each type of data. Once the training process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output. (Javapoint, 2011), Regression analysis will be performed using the car-prices dataset, which was obtained from the Kaggle website. The purpose of the regression analysis is to develop a model that can precisely forecast car prices while accounting for a variety of feature factors.

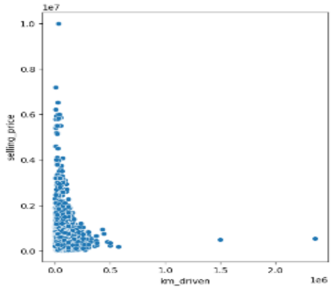
Existing Literature: “As the number of previous owners increases, the selling price of a vehicle tends to decrease." This can be partly attributed to the car age, as vehicles with more owners are likely to be older. Analysis also suggests that higher mileage translates to a lower selling price, indicating increased wear and tear on the vehicle. (Medium, 2012) Acknowledges fuel type as a factor in car price prediction models, implying its market significance. (Gate, 2008)

Research Questions

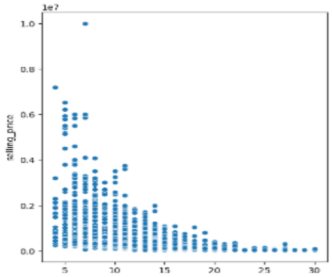
1. Explore the relationship variables ‘Car_age’ and ‘Km_driven’ have on the selling price of a car.
2. Is there a change in selling price of cars due to factors such as the fuel type and seller type?
3. Explore the change in car prices over the years .

Exploratory Data Analysis (EDA)

Exploratory data analysis was conducted in order to gain insight into the research questions identified above.

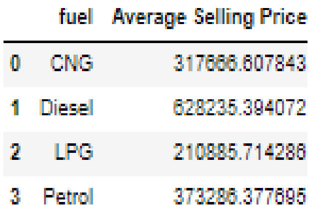


```
([[ 1.        , -0.16462831],
  [-0.16462831,  1.        ]])
```



```
([[ 1.        , -0.47687171],
  [-0.47687171,  1.        ]])
```

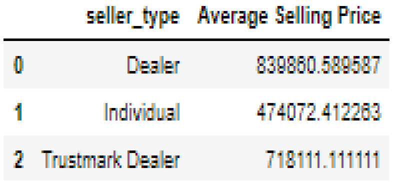
The scatterplots show if any linear relationship exists between the variables ‘Km_driven’ , ‘Car_Age’ with selling price of a car . The correlation matrix for each scatterplot is given under each scatterplot , it can be seen there is a weak negative correlation for the relationship between ‘Km_driven’ and selling price of a car which suggests Cars with higher mileage tend to sell for less, but other factors likely matter more. The correlation matrix shows a negative moderate relationship for the relationship between ‘Car_Age’ and the selling price of a car which suggests as car age increases, selling price tends to decrease, but other factors likely influence price as well.



	fuel	Average Selling Price
0	CNG	317666.607843
1	Diesel	628235.394072
2	LPG	210885.714286
3	Petrol	373286.377695

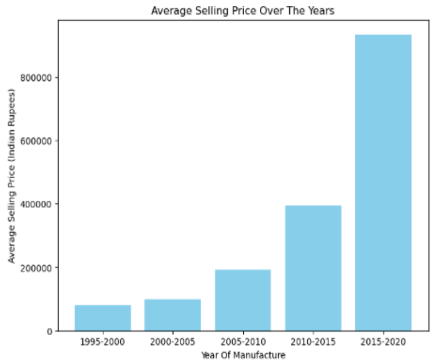
It can be seen that on average diesel cars are the most expensive vehicles whereas LPG vehicles are the cheapest vehicles . Therefore, we can see that fuel type does in fact have an impact on the selling price of a car , as we see different average prices for the different fuel type vehicles. The difference in prices for the fuel types may be justified with the benefits of the fuel type , eg: diesel -potentially higher performance

.



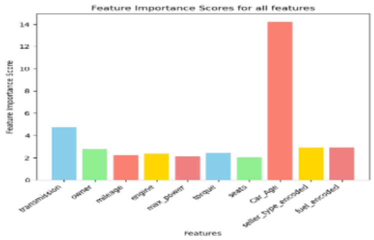
	seller_type	Average Selling Price
0	Dealer	839860.589587
1	Individual	474072.412263
2	Trustmark Dealer	718111.111111

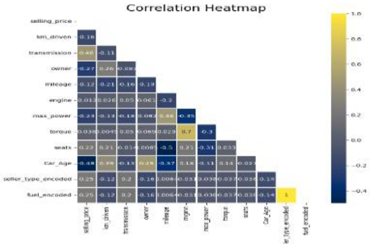
The average selling price of vehicles when sold by dealers is highest in comparison to the other seller types . the higher average price at dealerships could be due to additional services, convenience, and perceived value they offer to some buyers.



There has been an upward trend in the average selling prices over time, with newer cars fetching greater prices. This might be explained by things like inflation, technological developments, enhanced car performance and safety, and changes in consumer tastes toward newer models. In the used car market, the scarcity and rarity of newer cars may also play a role in their higher costs. All things considered, the evidence points to a favorable association between the average selling price of cars and the year of manufacturing.

Feature Importance/ Feature Selection

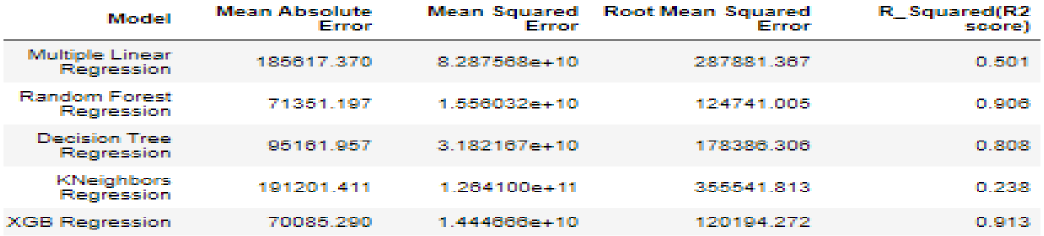




This bar plot shows us the feature importance score of all variables , the top 5 most important variables are 'Car_Age', 'transmission' , 'owner' ,' seller_type' and 'fuel' . 'Car_Age' is the most important variable for our analysis and the least important variable would be the number of seats in the car.

High correlation between predictor variables in a regression model is known as multicollinearity, and it can skew the interpretation of the effects of individual variables. However, we don't see high correlation between our predictor variables as the correlation between all features are low or moderate therefore, we will not need to drop any feature when fitting our regression models.

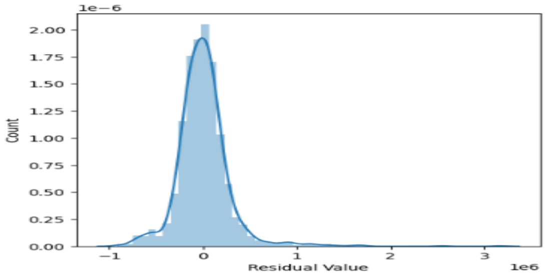
Evaluation Of The Regression Models

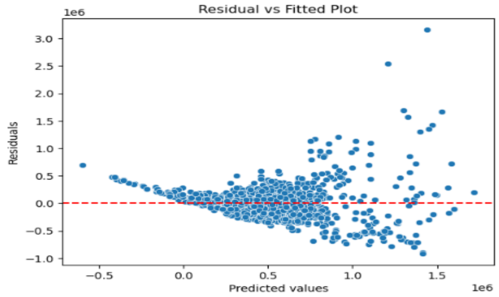


Model	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	R_Squared(R2 score)
Multiple Linear Regression	185617.370	8.287568e+10	287881.367	0.501
Random Forest Regression	71351.197	1.556032e+10	124741.005	0.906
Decision Tree Regression	95161.957	3.182167e+10	178386.306	0.808
KNeighbors Regression	191201.411	1.264100e+11	355541.813	0.238
XGB Regression	70085.290	1.444666e+10	120194.272	0.913

Multiple Linear Regression (MLR)

Using two or more input variables, multiple linear regression makes predictions about a single result. However, it is essential under a MLR approach that the basic assumptions of a linear regression model are satisfied. Such as the relationship between the independent and dependent variables is linear, observations are independent of each other ,the variance of the errors is constant across all levels of the independent variables, the residuals (errors) are normally distributed, the independent variables are not highly correlated with each other.





The distribution plot (left) shows the error terms appear to normally distributed (approximately) , therefore this assumption holds. The scatterplot (right) was plotted to visualize the residuals against the predicted values , The blue line's placement at zero on the figure indicates that there isn't much bias present in the errors. The homoscedasticity assumption, which denotes consistent variance of errors across all levels of the independent variables, is probably satisfied based on this alignment. As identified under feature selection (previously) there is no perfect multi-collinearity in the model as no predictor variables in the model had a perfect positive or perfect negative correlation (+1 or -1) therefore that assumption holds. The linearity assumption was assessed by plotting scatterplots and getting the Pearson correlation between each predictor variable in MLR model with the target variable 'selling _price'. By inspecting scatterplots and correlation score the assumption made regarding linearity can be held valid. In order to test the assumption for autocorrelation a Durbin-Watson test was conducted where the null and alternative hypothesis would be ,H0: There is no autocorrelation in the model & H1: There is autocorrelation in the model. The Durbin Watson test statistic as per calculations done is 1.747 which is less than 2 but closer to 2 , at 5 % significance level we fail to reject the null hypothesis and conclude that there is no autocorrelation in the model . Therefore, it can be seen all assumptions required to carry out MLR have been satisfied , MLR can be used for car price prediction. The MLR model shows moderate performance with relatively high error metrics and accounts for 50% of the variance.

Now let us evaluate the other regression models : Performance metric as shown in the table above R^2 score we would want to be as high as possible as this shows the variation in the model explained by the predictor variables , the metric such as MAE (mean absolute error),MSE(mean squared error) and RMSE(root mean squared error) we would want to be as low as possible as these represent measures of error and affect the accuracy of predictions .The Random Forest and XGB models outperform others with significantly lower errors and higher R^2 scores (90.6% for Random Forest and 91.3% for XGB), explaining most variance in the target variable. The Decision Tree performs reasonably well but not as effectively as Random Forest. K Neighbors demonstrates the weakest performance with higher errors and a low R^2 score (23.8%). Overall, Random Forest and XGB Regression models excel in accuracy and explanatory power.

Conclusion: As shown under EDA variables ‘Car_Age’ and ‘km_driven’ does have some impact on the car price, however not very significant , which suggests other factors play a vital role in price prediction. Also, it can be concluded that car prices do change due to factors such as fuel type and seller type due to the effective benefits (mentioned under EDA) and services provided by the seller type (mentioned under EDA). Car prices have also shown to increase over recent manufacturing years , this could be due to inflation, economic climate change , technological advancements etc.

Task : Supervised Learning - Classification

Introduction : Supervised learning classification involves training a model using labeled data to predict the category of new, unseen data . (Simplilearn, 2010).The dataset used for classification is on airplane customer satisfaction , this was downloaded from Kaggle website .Different classification models have been created to group customers as ‘satisfied’ or ‘dissatisfied’ based on available data .

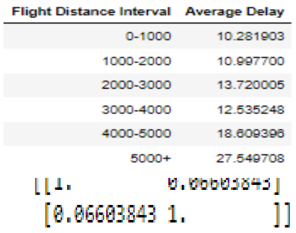
Existing Literature : “results imply that the type of delay cause can be distance dependent, suggesting a potential connection between flight distance and the likelihood of experiencing delays." (ResearchGate, 2008)"Airline service quality, particularly on-time performance, has a significant positive impact on passenger satisfaction."(Journal of Air Transport Management, 2009) “Airlines need to consider age when designing their service offerings, as different age groups prioritize various service elements and have varying levels of satisfaction." (Journal of Air Transport Management, 2017) “results suggest that service quality has a significant positive effect on customer satisfaction, which in turn, has a significant positive effect on airline loyalty." (Journal of Travel Research, 2012)

Research Questions

1. Do longer flights cause more delays for customers ?
2. How much of an impact does flight delay have on customer satisfaction?
3. How does customer satisfaction vary across all customer age groups?
4. Explore the relationship between customer loyalty and customer satisfaction.

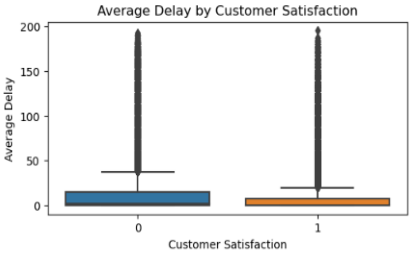
Exploratory Data Analysis

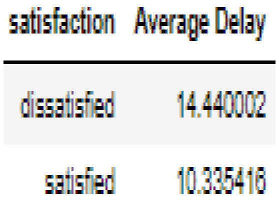
Exploratory data analysis was conducted to gain deep understanding and insight into research aims as implied by the research questions stated above. As per the first research requirement , the table (left) shows the average delay faced at different flight distance intervals (km). We can see overall as the flight distance the average delay faced increases , however it should be noted that the delay faced 2000-3000 range is higher than the 3000-2000. This could be due to factors like Air traffic congestion , airport infrastructure etc. The correlation between ‘Flight Distance’ and ‘Average Delay’ shows a weak positive relationship which indicates that even though there is some relationship between the variables it isn’t significant.



Flight Distance Interval	Average Delay
0-1000	10.281903
1000-2000	10.997700
2000-3000	13.720005
3000-4000	12.535248
4000-5000	18.609396
5000+	27.549708

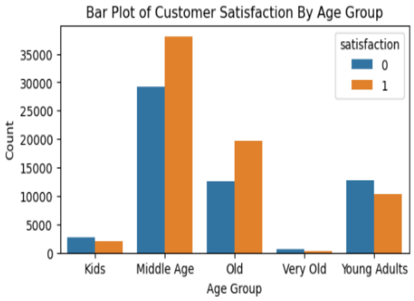
```
[[1.         0.06603843]
 [0.06603843 1.        ]]
```



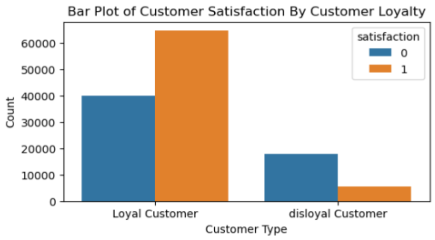


satisfaction	Average Delay
dissatisfied	14.440002
satisfied	10.335416

It can be seen dissatisfied customers appear to have faced higher delay on average when compared to satisfied customers, but the difference is not very significant as shown in the table and boxplot on the left . The point biserial correlation score between ‘satisfaction’ and ‘Average Delay’ is -0.07914 , which shows a weak negative correlation indicating that higher the delay faced lower the satisfaction of customers, but it does also very much indicate that there are other factors that contribute to customer satisfaction besides delay faced by customers.

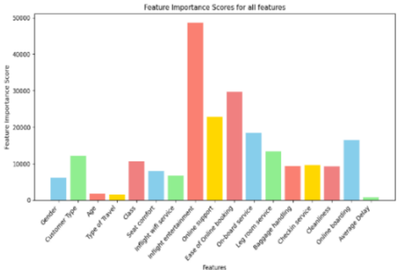


When assessing customer satisfaction levels across age groups 42.71% of all kids who flew are satisfied , 44.8% of all young adults who flew are satisfied , 56.7% of all middle age consumers who flew are satisfied 61.07% of all old age consumers who flew are satisfied and 30.1% of very old consumers who flew are satisfied , however the statistics for 'Kids' , 'Young Adults' , ' Very Old' aren’t promising requires further research as these age groups have higher dissatisfied customers than those satisfied.



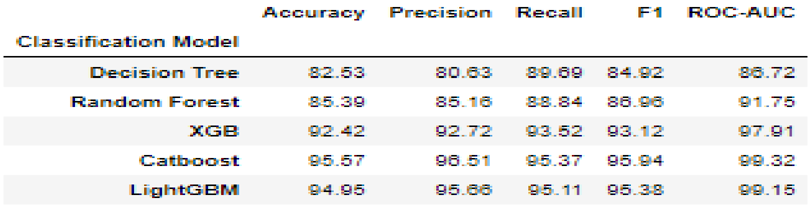
When exploring the relationship between customer loyalty and customer satisfaction it can be seen 61.8% of loyal customers are satisfied and 24% of disloyal customers are satisfied . But having disloyal customers who are satisfied is promising , the disloyalty can be from them being less informed about the airline services and could be a marketing problem. It can also be seen 38.2% of loyal customers are dissatisfied which is alarming, therefore research may be required to understand and identify reasons for customer dissatisfaction. Also, of disloyal customers 76% of them were dissatisfied , it may help to learn from these customers the reasons for their dissatisfaction when making improvements to the airline service.

Feature Importance/ Feature Selection



‘Inflight entertainment’ has the highest feature score and ‘Average Delay’ the lowest feature score. When selecting features for the model 'Age group', ‘Departure Delay in Minutes' and 'Arrival Delay in Minutes’ were dropped as these variables were represented under features ‘Age’ and ’Average Delay’. ‘Flight Distance’ was also dropped as its feature score was low in comparison to other features. ‘Seat comfort’ and ‘Food and drink’ showed a high correlation of 0.71 , therefore in order to avoid multi collinearity ‘Seat comfort' was dropped as it has a lower correlation with the target variable. Correlations among other variables appeared to be moderate and therefore selected for model building.

Evaluation Of Classification Models



Classification Model	Accuracy	Precision	Recall	F1	ROC-AUC
Decision Tree	82.53	80.63	89.69	84.92	86.72
Random Forest	85.39	85.16	88.84	86.96	91.75
XGB	92.42	92.72	93.52	93.12	97.91
Catboost	95.57	96.51	95.37	95.94	99.32
LightGBM	94.95	95.66	95.11	95.38	99.15

These classification models cover a range of effective techniques. Decision Trees offer simplicity, Random Forests combine many trees for better accuracy, and XGBoost and LightGBM use boosting methods for high performance. Catboost specializes in handling categorical data well. Together, they make a versatile set of tools for different classification tasks. Accuracy gauges overall correctness, while the F1 score strikes a balance between precision and recall, crucial for handling class imbalances. Recall ensures all positive instances are captured, crucial for avoiding missed detections. Precision emphasizes minimizing false positives, essential in scenarios with stringent cost considerations. Lastly, the ROC-AUC score offers a holistic view of the model's discrimination abilities across different thresholds. Class balance in the target variable is crucial for interpreting accuracy because it ensures that the model is not biased towards predicting the majority class. However, in the dataset used classes are not heavily imbalanced (55:45) , accuracy can still provide a reliable measure of overall model performance .

The Catboost classifier stands out as the top performer among the evaluated models, boasting the highest accuracy, precision, recall, F1 score, and ROC-AUC score. Its exceptional performance across all metrics indicates its robustness in accurately predicting the target variable. Following closely, the XGB and LGBM models also demonstrate strong performance across all metrics, showcasing the effectiveness of gradient boosting techniques in classification tasks. Random Forest achieves respectable results with balanced accuracy, precision, recall, and F1 score, suggesting its suitability for handling complex datasets. While the Decision Tree model lags slightly behind the others in terms of accuracy and F1 score, it still demonstrates reasonable performance, particularly in recall. Overall, these results underscore the effectiveness of ensemble and boosting algorithms, with Catboost emerging as the top choice for achieving high predictive performance in classification tasks.

Conclusion

Previous and exploratory research suggests that while flight distance does have some impact on delays, it's not significant compared to other factors. Flight delays do affect customer satisfaction, but the difference in delay between satisfied and dissatisfied customers isn't substantial, though dissatisfied customers tend to experience higher delays. Additionally, dissatisfaction levels vary across age groups, with younger and older customers showing higher dissatisfaction, indicating a need for airlines to reassess their services. Interestingly, customer loyalty doesn't guarantee satisfaction, as a significant portion of loyal customers are dissatisfied, suggesting loyalty is influenced by various factors beyond satisfaction.

Bibliography

Builtin. (2022, March 13). *step-step-explanation-principal-component-analysis*. Retrieved from builtin: https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Gate, R. (2008, March 15). *Car price prediction using machine learning techniques*. Retrieved from Research Gate: https://www.researchgate.net/publication/331994496_Car_price_prediction_using_machine_learning_techniques

IBM. (1995, March 11). *Unsupervised Learning*. Retrieved from IBM: https://www.ibm.com/topics/unsupervised-learning

Javapoint. (2011, March 15). *Javapoint*. Retrieved from Supervised Machine Learning: https://www.javatpoint.com/supervised-machine-learning

JavaPoint. (2011, March 13). *k means clustering algorithm in machine learning*. Retrieved from JavaPoint: https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

Medium. (2012, March 14). *Statistical Analysis Used Car Price*. Retrieved from Medium: https://medium.com/@dsyafz/statistical-analysis-used-car-price-132b073439d5

ResearchGate. (2008, March 16). *Generation and prediction of flight delays in air transport*. Retrieved from ResearchGate: https://www.researchgate.net/publication/351474946_Generation_and_prediction_of_flight_delays_in_air_transport

Simplilearn. (2010, March 18). *learn machine learning-algorithms free course skillup*. Retrieved from Simplilearn.com: https://www.simplilearn.com/learn-machine-learning-algorithms-free-course-skillup