ST3189 - Course work

Nikhila Rayalacheruvu

Dataset 1:

In this project, I aim to analyze online shoppers' behavior using the Online Shoppers Purchasing Intention Dataset. The dataset contains anonymized information collected from an online retail store's website and includes various attributes describing visitors' behavior and characteristics. My goal is to build predictive models to identify potential customers who are likely to make a purchase based on their browsing behavior and other attributes.

- **Source**: UCI Machine Learning Repository
- **Format**: CSV (Comma Separated Values), No NA values
- **Size**: 12330 rows, 18 attributes
- **Attributes**: Attributes list is added in References as it is acquired from Research papers. [4] . They contain session information like duration and demographic information like traffic source,Operating system etc.,

Research Questions:

1. How can e-commerce businesses leverage machine learning to optimize marketing strategies?
2. Can I accurately predict whether a visitor will make a purchase based on their behavior and demographics?
    a. What are the key factors influencing online purchase intent?
    b. How do different machine learning algorithms compare in predicting online purchase intent?

Research Question1:

Understanding the behavior and preferences of online shoppers is crucial for businesses to effectively target and engage their audience. In this study, I explore segmentation analysis of online shoppers using K-means clustering and investigate how businesses can optimize their marketing strategies accordingly.

Methodology:

Since K means clustering requires its variables to be continuous , time related variables in Dataset-1 are used for this analysis ("PageValues", "BounceRates", "Administrative_Duration", "ProductRelated_Duration", "Informational_Duration"). Through K-means clustering, I segmented the online shoppers into 3 distinct groups

1. **High Engagement :**
   Users in green clusters (Fig:1.1) who spend a lot of time on the website either browsing admin pages , product related pages or information pages.
2. **Normal Users:**
   Users in black clusters (Fig:1.1) who spend some time on the website either browsing admin pages , product related pages or information pages.
3. **High Bounce rate :**
   Users in red clusters (Fig:1.1) who bounce out of pages a lot.

These clusters are then applied to the original data set to analyze behaviors of different clusters

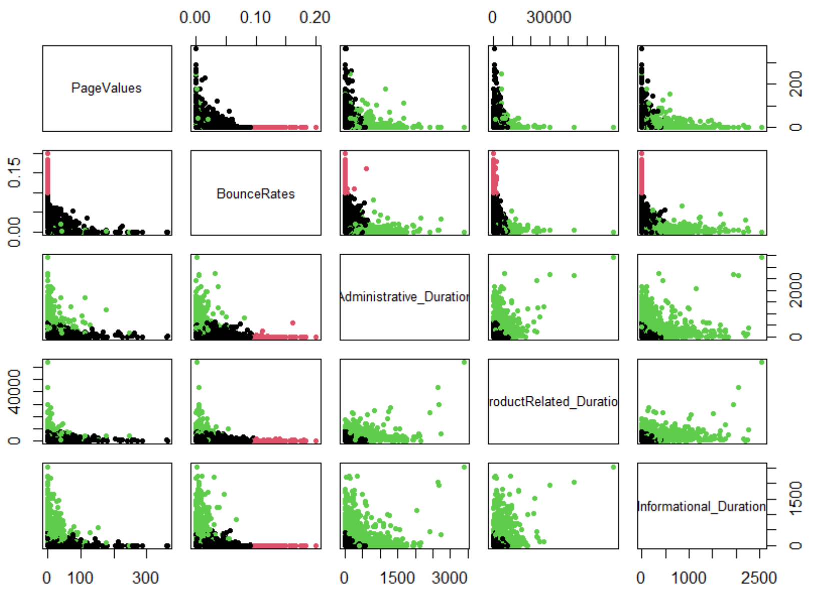


Fig1.1: Represents scatter plot of Clusters across the model variables.

Insights:

1. Surprisingly, High engagement sessions show a preference for pages with lower page values. This might mean that the users follow an exploratory pattern and the conversion rate of time to revenue is much lower than a normal intent driven user. (Fig1.1:)
2. There is a clear correlation between bounce rate and page value. Users with higher bounce rates are more likely to visit pages with lower page values, indicating a potential mismatch between user expectations and the content provided on these pages.(Fig1.1:)
3. Traffic type 2 emerges as a significant driver of user engagement (Fig1.2:)
4. New users have lesser bounce rate than Returning Users , hence it shows that new users show exploratory behavior pattern.(Fig1.3:)

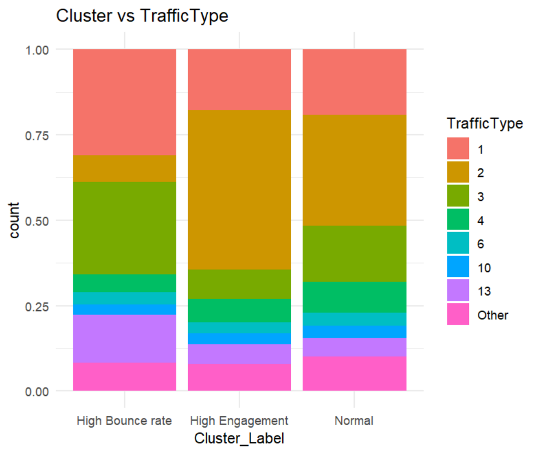


Fig1.2: K means cluster vs Traffic type

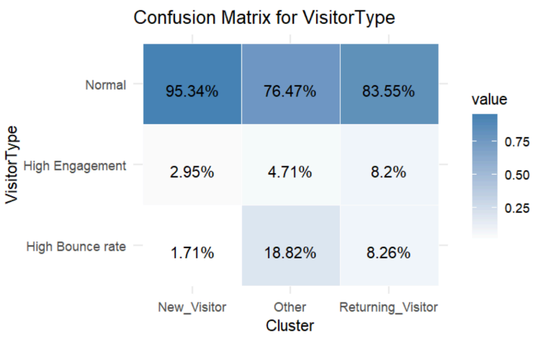


Fig1.3: K means cluster vs Visitor type

Conclusion:

Tailoring marketing strategies to target each shopper segment effectively is crucial for maximizing conversions and optimizing marketing ROI. Based on the segmentation analysis, businesses can implement the following strategies:

- **Engagement Enhancement:** Implement strategies to enhance engagement and reduce bounce rates for high bounce sessions to encourage them to explore further and make purchases.
- **Conversion Optimization:** By identifying high engagement shoppers I can target impulse buy categories like fashion towards them and increase conversion rates
- **Traffic Optimization:** Allocate marketing resources effectively by focusing on channels that drive the highest engagement rates, such as Traffic Type 2.

Also by revising these segments regularly and revisiting the analysis will help the company create fresh real time perspective on the users.

Research Question2:

To be able to answer these questions I leverage machine learning models like Logistic regression , CART and XGboost. The "Revenue" column in our dataset indicates whether a session resulted in a purchase, serving as our target variable for prediction. The data set has 12330 entries with only 1908 true revenue, showing that the data set is heavily imbalanced.

With 18 features in the dataset, indicating a multinomial equation, feature selection becomes pivotal for model building. I embark on an iterative process to identify significant predictors.

Literature Review:

There are many studies which compared the different machine learning models to identify the purchase intent of the user. Here I show the different literatures that use the same dataset.

[3] This is the original paper on this database. The paper explores two methods of identifying the purchase intention , one with help of clickstream page view data and applied oversampling and feature selection methods. The Multilayer Perceptron Network (MLP) model showed an accuracy of 87% and F1 score of 0.86. In the other method they used a Long Short-Term Memory-based Recurrent Neural Network (LSTM-RNN) to identify users who are moving out of the web page without purchasing. The final objective is to find out users for whom they can offer content.

[2] This paper acts as an extension and proposes a need for a model that forecasts the visitor behavior before the start of the session. It concludes that with an accuracy of 86% the Random forests classifier is able to fit the problem statement.

Methodology:

Exploratory Data Analysis:

Figures *2.1, 2.2, 2.3* explore the relationship between page value , administrative and exit rates with revenue. The box plots show that the revenue true and revenue false have different values of selected variables.

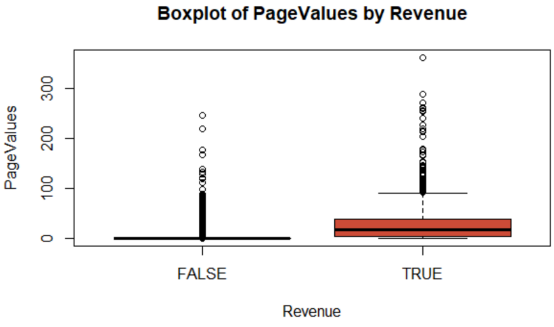


Fig2.1: Box plot of page values vs revenue

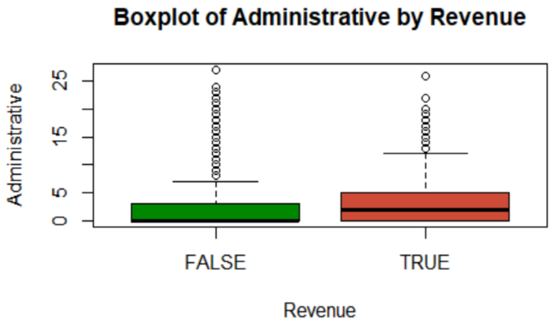


Fig2.2: Box plot of page values vs revenue

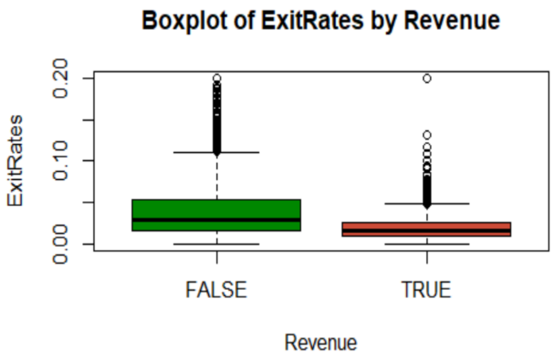
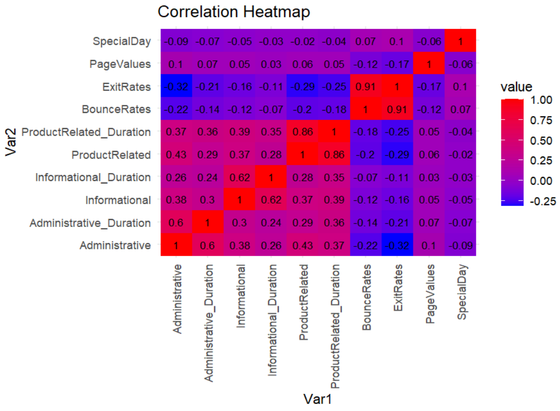


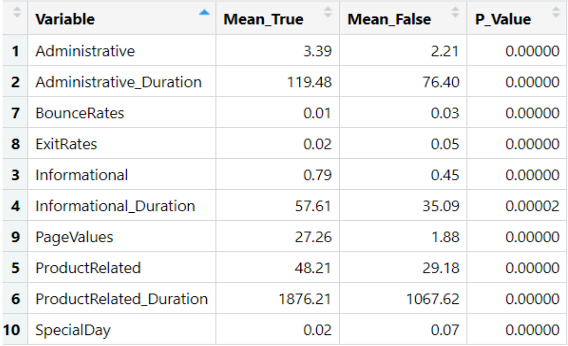
Fig2.2: Correlation between All numeric variables



Var2										
SpecialDay	-0.09	-0.07	-0.05	-0.03	-0.02	-0.04	0.07	0.1	-0.06	1
PageValues	0.1	0.07	0.05	0.03	0.06	0.05	-0.12	-0.17	1	-0.06
ExitRates	-0.32	-0.21	-0.16	-0.11	-0.29	-0.25	0.91	1	-0.17	0.1
BounceRates	-0.22	-0.14	-0.12	-0.07	-0.2	-0.18	1	0.91	-0.12	0.07
ProductRelated_Duration	0.37	0.36	0.39	0.35	0.86	1	-0.18	-0.25	0.05	-0.04
ProductRelated	0.43	0.29	0.37	0.28	1	0.86	-0.2	-0.29	0.06	-0.02
Informational_Duration	0.26	0.24	0.62	1	0.28	0.35	-0.07	-0.11	0.03	-0.03
Informational	0.38	0.3	1	0.62	0.37	0.39	-0.12	-0.16	0.05	-0.05
Administrative_Duration	0.6	1	0.3	0.24	0.29	0.36	-0.14	-0.21	0.07	-0.07
Administrative	1	0.6	0.38	0.26	0.43	0.37	-0.22	-0.32	0.1	-0.09
Var1	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay

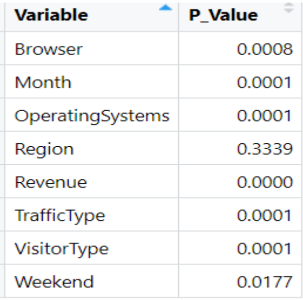
Fig2.4: Correlation between All numeric variables

I also explore correlations among numerical features. Notably, "Bounce rate" and "Exit rate" exhibit a high correlation of 0.91, while "Product related duration" and "Product related" also demonstrate significant correlation. Thus, I must carefully decide which features to retain.



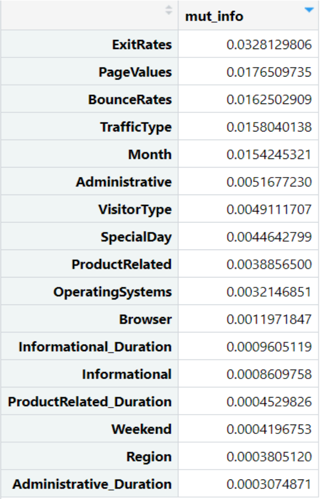
	Variable	Mean_True	Mean_False	P_Value
1	Administrative	3.39	2.21	0.00000
2	Administrative_Duration	119.48	76.40	0.00000
7	BounceRates	0.01	0.03	0.00000
8	ExitRates	0.02	0.05	0.00000
3	Informational	0.79	0.45	0.00000
4	Informational_Duration	57.61	35.09	0.00002
9	PageValues	27.26	1.88	0.00000
5	ProductRelated	48.21	29.18	0.00000
6	ProductRelated_Duration	1876.21	1067.62	0.00000
10	SpecialDay	0.02	0.07	0.00000

Fig2.5: P value of Numeric variables using means



Variable	P_Value
Browser	0.0008
Month	0.0001
OperatingSystems	0.0001
Region	0.3339
Revenue	0.0000
TrafficType	0.0001
VisitorType	0.0001
Weekend	0.0177

Fig2.6: P value of Categorical variables using Fisher test



	mut_info
ExitRates	0.0328129806
PageValues	0.0176509735
BounceRates	0.0162502909
TrafficType	0.0158040138
Month	0.0154245321
Administrative	0.0051677230
VisitorType	0.0049111707
SpecialDay	0.0044642799
ProductRelated	0.0038856500
OperatingSystems	0.0032146851
Browser	0.0011971847
Informational_Duration	0.0009605119
Informational	0.0008609758
ProductRelated_Duration	0.0004529826
Weekend	0.0004196753
Region	0.0003805120
Administrative_Duration	0.0003074871

Fig2.7: Mutual information of all variables

Feature Selection:

Relationship between all available variables and Revenue is calculated using p values and Fisher Test. Region seems to be variable with very low significance. (Fig2.6). Following the methodology outlined in [1] I prioritize the identification of mutual information, followed by the application of mutual relevance and mutual redundancy concepts. Mutual Information calculation requires all variables to be converted into categories , and hence i bucketed all numeric variables into buckets of 5. Based on this approach, I opted to remove features such as "Informational_Duration," "Informational," "BounceRates," "ProductRelated_Duration," and "Administrative_Duration," as they exhibit lower mutual information compared to their correlated counterparts. (Fig2.7:)

Initial iteration of logistic model is performed to identify low significance Variables. Features of Browser, Operating system and Region are removed. Weekend though seems a relevant factor for revenue , does not contribute to the calculation of revenue in presence of other variables

Model Building :

Using machine learning algorithms, including Logistic Regression, Classification and Regression Trees (CART), and XGBoost, I constructed predictive models trained on our dataset. I undersampled the data for some models to address class imbalance:

Base Model with Logistic Regression: Logistic regression is used as a baseline model to establish a starting point for predicting purchase intent based on customer behavior and demographic information. It provides a straightforward binary classification of whether a visitor is likely to make a purchase or not.

Train-Test Split: Then the dataset was divided into 70% for training and 30% for testing to evaluate model performance on unseen data.

CART Decision Tree: Classification and Regression Trees are flexible, easy to interpret, and can be highly accurate and stable. Decision trees solve problems by splitting data, based on certain variables, to identify the best path to make an accurate prediction

I applied a CART decision tree algorithm, selecting a tree with 9 splits as 1 split was too simplistic for our dataset.

Random Forest with Cross-Validation: Random Forest is utilized to improve prediction accuracy by aggregating multiple decision trees. Random forests are generally considered as a predictor that outperforms many other classification algorithms.[2]

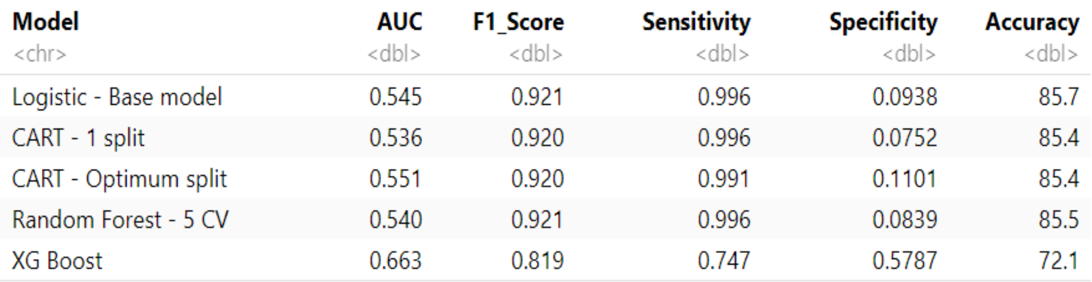
I performed Random Forest with 5-fold cross-validation to account for model variability and optimize generalization to unseen data.

XGBoost: XGBoost is employed to further enhance prediction performance by sequentially building a series of weak learners and optimizing model parameters to minimize prediction errors. The model's ability to handle nonlinear relationships and large datasets makes it effective in capturing subtle patterns in customer behavior and making accurate predictions of purchase intent.

Employing XGBoost, I identified an optimal cutoff using the AUC-ROC curve to balance sensitivity and specificity.

Performance Evaluation:

Comparison of different models before Under sampling:



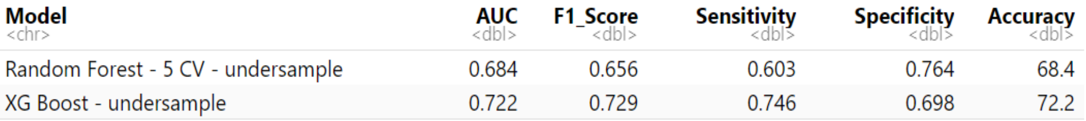
Model	AUC	F1_Score	Sensitivity	Specificity	Accuracy
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Logistic - Base model	0.545	0.921	0.996	0.0938	85.7
CART - 1 split	0.536	0.920	0.996	0.0752	85.4
CART - Optimum split	0.551	0.920	0.991	0.1101	85.4
Random Forest - 5 CV	0.540	0.921	0.996	0.0839	85.5
XG Boost	0.663	0.819	0.747	0.5787	72.1

So far based on the unbalanced data set , due to the prevalence of negative samples, our models tend to exhibit high sensitivity but suffer from low specificity. This imbalance results in the misclassification of many positive revenue conversions as negative, leading to missed revenue opportunities.

For instance, the base logistic model demonstrates a specificity of 0.09, indicating its propensity to misclassify positive cases. Similarly, even Random Forest, often touted for its efficacy with skewed datasets, yields a specificity of 0.08. However, XGBoost, while reducing sensitivity, notably improves specificity to 0.57. And , XGBoost boasts a superior AUC compared to other models, indicating its overall better performance in distinguishing between positive and negative cases.

To address the skewed nature of the dataset, we employed undersampling, a technique aimed at creating a more balanced representation of both positive and negative classes. While this approach may result in some loss of information, it helps mitigate the imbalance issue and enables our models to better capture the underlying patterns within the data.

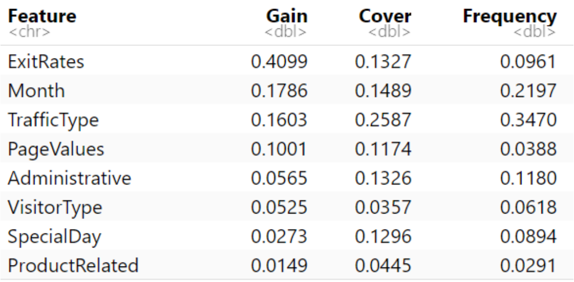
Comparison of different models after Under sampling:



Model	AUC	F1_Score	Sensitivity	Specificity	Accuracy
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Random Forest - 5 CV - undersample	0.684	0.656	0.603	0.764	68.4
XG Boost - undersample	0.722	0.729	0.746	0.698	72.2

There is a significant increase in the specificity of the Random forest model but still falls behind XG boost in F score, AUC and accuracy metrics.

Variable Importance:



Feature	Gain	Cover	Frequency
<chr>	<dbl>	<dbl>	<dbl>
ExitRates	0.4099	0.1327	0.0961
Month	0.1786	0.1489	0.2197
TrafficType	0.1603	0.2587	0.3470
PageValues	0.1001	0.1174	0.0388
Administrative	0.0565	0.1326	0.1180
VisitorType	0.0525	0.0357	0.0618
SpecialDay	0.0273	0.1296	0.0894
ProductRelated	0.0149	0.0445	0.0291

This is the result of Variable importance in XGBoost - undersample model.
Feature: The name of the feature.
Gain: The improvement in accuracy resulting from a split based on this feature.
Cover: The relative quantity of observations concerned by a feature.
Frequency: The relative frequency of a feature.

It clearly shows that Exit rates are a major contributor to identify the revenue status though it has lower frequency.

Conclusion:

Our experimental results revealed the following key findings:

1. Overall, ensemble methods like Random Forest and XGBoost showed promising results for predicting online shoppers' purchasing intention, with undersampling techniques further enhancing their performance.
2. Exit rates and Traffic type are important factors in determining the possible revenue during a session. Month and Traffic type are the factors that we can influence and use them for improving marketing strategies.

Further experimentation and optimization could lead to even better performance and insights.

Data Set 2 :

Introduction:

In this study, I aim to evaluate various variable selection and dimensionality reduction techniques for predicting the sale price of houses using a curated dataset obtained from the Assessor's Office in Ames. Our primary objective is to identify the most effective machine learning methodologies that yield accurate predictions of the sale price.

To achieve this, we employ different variable selection and dimensionality reduction methods and assess their performance in predicting the sale price, which serves as our target variable (Y). The dataset undergoes preprocessing to ensure cleanliness and relevance to the task at hand.

- **Format**: ARFF file , No NA values
- **Size**: 2930 rows, 81 attributes
- **Attribute Link**: Since the data set is huge , the list of attributes are mentioned Here.

As an overview the data set contains multiple features like Building_type, Neighbourhood, Areas of different places in the house, central_air etc.,

Research Question3:

1. What are the most important factors affecting housing prices ?
2. How do different machine learning methods compare in high dimensional data like this data set?

Methodology:

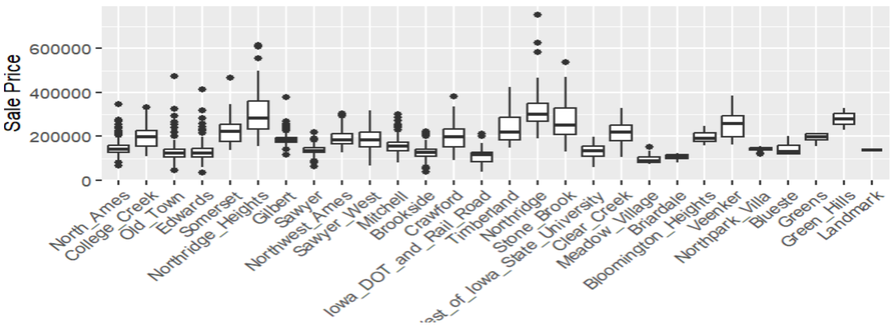
Data Cleaning & Feature Selection:

- **Identification of Categorical Variables**: Among the 81 features, I identified 46 as categorical variables with more than 2 levels and converted ordinal categorical features like Bsmt_Cond, Bsmt_Qual, Kitchen_Qual, Overall_Cond, and Land_Slope into numerical columns, respecting their ordinal nature.

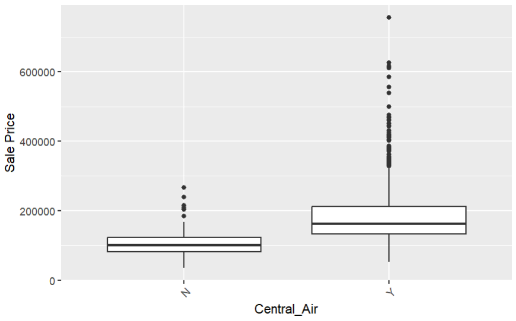
- **Handling Nominal Variables**: I scrutinized nominal variables, grouping together levels with minimal statistical significance. This step ensured that my dataset remained concise and focused on meaningful variations.
- **Correlation Analysis**: I examined numeric features with pairwise correlations exceeding 0.75. Those showing lower correlation with Sale_price compared to their counterparts were pruned from the dataset. For instance, variables like Garage_Area, Fireplaces, and Pool_Area, while correlated with other features, exhibited limited influence on Sale_price.
- **Visualization and Variation Analysis**: I visually inspected each categorical variable through box plots against Sale_price. This analysis helped me identify variables with low variation across Sale_price, leading to their removal or the combination of levels. This ensured that only impactful categorical variables were retained for further analysis.
- **Consideration of Sale Conditions**: I scrutinized the Sale_condition variable, with its various levels such as Family and Partial. Recognizing that different sale conditions may impact pricing differently, I chose to consider only the "Normal" condition for subsequent analysis, ensuring a more focused examination.
- **Log Transformation Evaluation**: While considering Sale_price, I evaluated log transformations due to its extensive range. However, applying logarithms did not yield a significant improvement in the R-square value. As a result, I retained the original Sale_price values for further analysis.

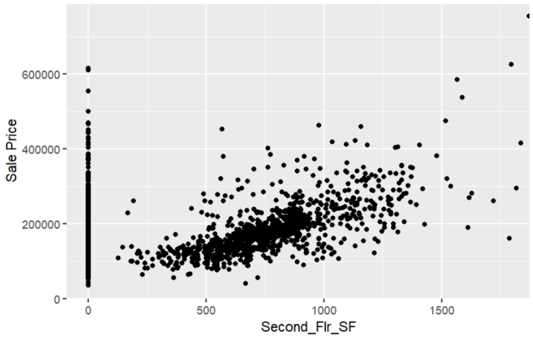
After this cleanup , we ended up with 68 features , 21 of them categorical.

Exploratory Data Analysis:



Neighborhoods like Veenker, North ridge and Stonebrook seem to be expensive places to buy homes.



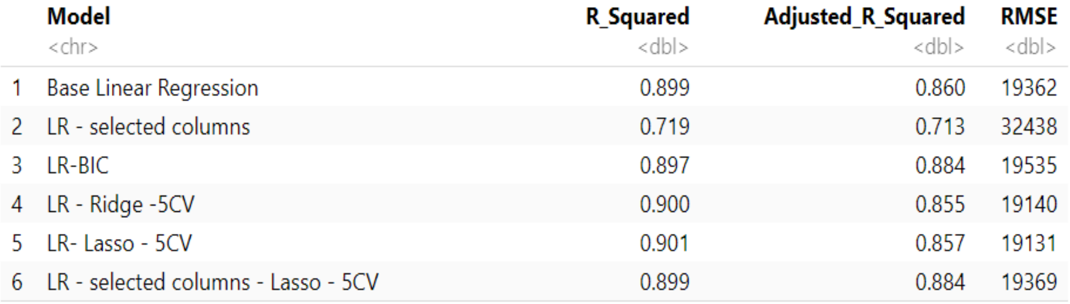


1. Houses with central air installed go at higher price than regular houses
2. Many Area related factors like basement_sf, loft_sf show high positive correlation with Sale_price. We can answer 70% of data variability with area factors alone.

Model Comparison:

8 different Models are compared here , Base linear regression utilizes all cleaned features. LR - selected columns are created for most straight forward associations with Sale_price. LR - BIC model uses step wise Bayesian information Criterion to filter The 68 features and identify optimum number of features. The optimum number is found to be 20. Ridge and Lasso are regularization techniques that penalize large coefficients in regression models. Lasso additionally introduces sparsity by encouraging some coefficients to be exactly zero. Cross validation is employed for us to identify the top performing hyper parameters.

The Models are trained on 70% data and the remaining is used for testing and thus calculating the below metrics.

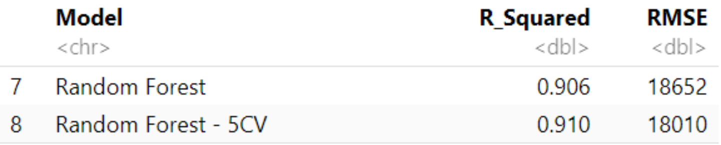


	Model	R_Squared	Adjusted_R_Squared	RMSE
	<chr>	<dbl>	<dbl>	<dbl>
1	Base Linear Regression	0.899	0.860	19362
2	LR - selected columns	0.719	0.713	32438
3	LR-BIC	0.897	0.884	19535
4	LR - Ridge -5CV	0.900	0.855	19140
5	LR- Lasso - 5CV	0.901	0.857	19131
6	LR - selected columns - Lasso - 5CV	0.899	0.884	19369

The models were evaluated based on their Adjusted R-squared and Root Mean Squared Error (RMSE) metrics. The base model achieved an R-squared of 0.899, indicating that it can explain 89.9% of the variability in the test data. However, it yielded an RMSE of $19,362, which, although not the highest, suggests room for improvement.

The model with selected columns, which utilized only 12 variables, showed a higher RMSE of $32,438, indicating a loss of information due to the random removal of variables. To address this, we employed the Bayesian Information Criterion (BIC) to select 20 variables with minimal loss.The best-performing model, Lasso Regression, utilized all 68 variables and achieved superior predictive performance. Despite its complexity, it leveraged the full spectrum of available information, resulting in a more accurate representation of the data.These results suggest that while both regularization techniques contribute to better predictive accuracy compared to the base model.

Adjusted R squared is lower for the Regularized models than the base model because of the high penalty parameter applied to them. This means that the model is not a very big improvement from the base model.

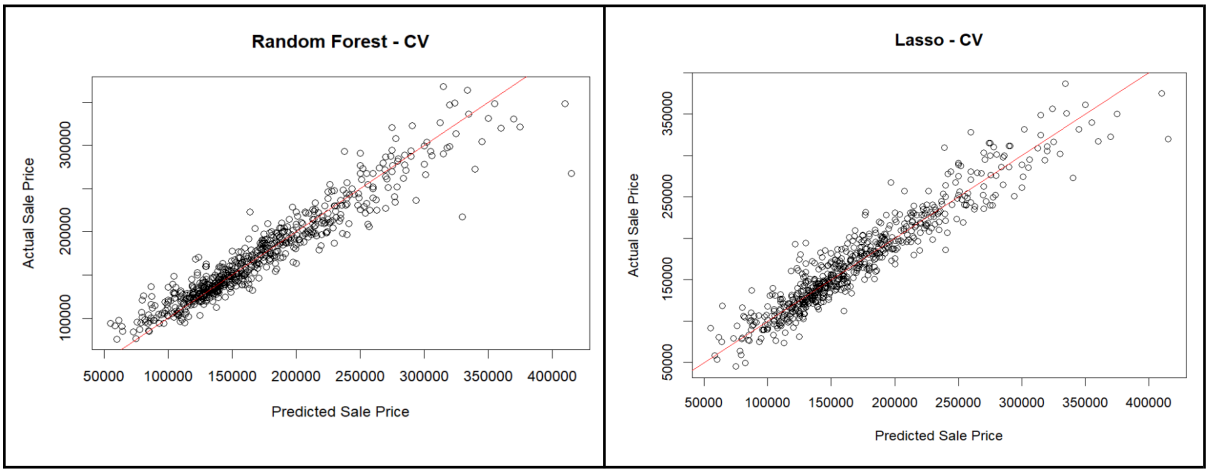


	Model	R_Squared	RMSE
	<chr>	<dbl>	<dbl>
7	Random Forest	0.906	18652
8	Random Forest - 5CV	0.910	18010

I also deployed Random Forests to compare with the regularization techniques, Though the Rsquared didn't improve significantly , the random forests decreased the RMSE significantly. Now i Compare the regression lines of Random forest cv and Lasso cv.

Predicted values Analysis:

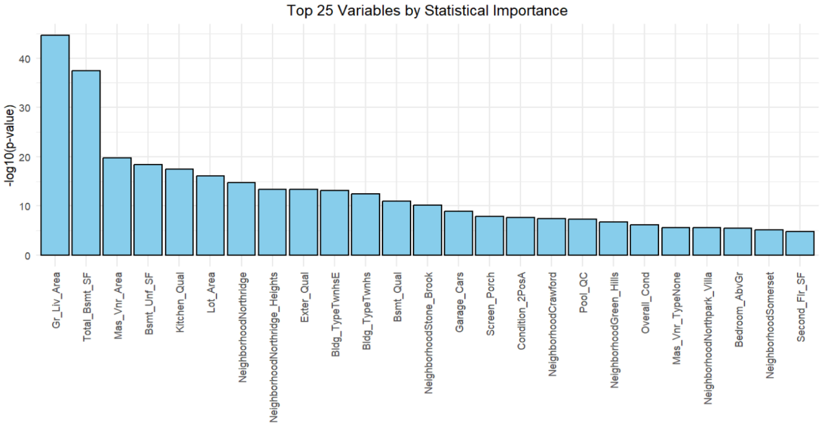
In both the models the Model line passes through most of the points other than the ones with



largest values , The above graphs also shows that our models are over estimating the sales values , especially the Random Forest model. The above graph also shows that Lasso - CV model is more near to the reality than the Random forest model. Hence our final model is Lasso - cv.

Conclusion:

Based on Variable Importance we can conclude the following:
I am using the Base model to calculate variable importance as it is more interpretable than lasso and Ridge.



We can see that The top variables influencing are Area size and neighborhood . Quality of the house comes after the area and neighborhood.

When looking at Quality Kitchen Quality is the most important factor people look at and Having a basement increases the price of the House considerably. Base ment increases the price more than a pool addition.
As Expected Town Houses are at higher price than other building types.

From the Model analysis we know that the Lasso - CV model works best for this scenario with RMSE $19131. Any regularization method generally performs better than the base model. For further improvement in RMSE I should try polynomials of these variables and interaction variables.

References:

[1] Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005 Aug;27(8):1226-38. doi: 10.1109/TPAMI.2005.159. PMID: 16119262.[PubMed]

[2] Baati K, Mohsil M. Real-Time Prediction of Online Shoppers' Purchasing Intention Using Random Forest. Artificial Intelligence Applications and Innovations. 2020 May 6;583:43–51. doi: 10.1007/978-3-030-49161-1_4. PMCID: PMC7256375.[ncbi]

[3] Sakar CO, Polat SO, Katircioglu M, Kastro Y. Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. Neural Comput. Appl. 2019;31(10):6893–6908. doi: 10.1007/s00521-018-3523-0. [Google Scholar]

[4]

Numeric Variables	
Administrative	Number of pages visited by the visitor in the "Administrative" category.
Administrative_Duration	Total time spent by the visitor on pages in the "Administrative" category.
Informational	Number of pages visited by the visitor in the "Informational" category.
Informational_Duration	Total time spent by the visitor on pages in the "Informational" category.
ProductRelated	Number of pages visited by the visitor in the "ProductRelated" category.
ProductRelated_Duration	Total time spent by the visitor on pages in the "ProductRelated" category.
BounceRates	Bounce rate of the webpage.
ExitRates	Exit rate of the webpage.
PageValues	Average value for a web page that a user visited before a transaction
SpecialDay	Indicates the closeness of the site visiting time to a special day.
Categorical Variables	
Month	Month of the year.
OperatingSystems	Operating system of the visitor.
Browser	Browser of the visitor.
Region	Geographic region from which the session has been started by the visitor.
TrafficType	Traffic source type.
VisitorType	Visitor type as returning visitor, new visitor, or other.
Weekend	Indicates whether the session is on a weekend or not.
Revenue	Indicates whether the visit resulted in a transaction or not.