DBA3803 Assignment 4 (E0200936)

1 Introduction

This assignment report attempts to break down and analyze the provided Bike Sharing dataset. The main objective is to obtain a better understanding of the user traffic of Bike Sharing Services, to accurately forecast bike usage, and to identify potential user segments to allow the Bike Sharing businesses to accurately target and attract more potential bike users. This is done through Regression, Classification and Clustering Analysis as elaborated in this report.

2 Regression Model

For the regression model, the target variable is *total_users*, which is derived from totaling number of casual and registered users at each hour. Apparent temperature was removed before regression, due to its high correlation with temperature (corr_coeff = 0.99). For regression, a total of 6 models were built. They include, Linear (with and without further variable selection), Lasso, Ridge, Elastic Net, and Random Forest Regression. The Random Forest Regression (RFR) Model was also built using Grid Search and 5-Fold Cross Validation (CV) to identify the optimal hyperparameters for the RFR Model with the best predictive power. Using 30% of the dataset as test set for out-of-sample cross validation, we can see that further feature selection, Lasso, Ridge and Elastic Net Regularizations improve the predictive performance (RMSE) of Linear Regression. This suggests that the optimal feature set should be smaller, which also leads to model parsimony while improving predictive power through reduced variance from the bias-variance tradeoff. Overall, as seen from the table on the left, the RFR model performs twice as well as other models and hence, is more ideal for predicting forecasts of future bike usage. However, RFR also has a slightly greater potential than other regression models to overfit, which may impede its ability to forecast future demand.

	RMSE
OLS (Variable Selection)	100.539048
Linear Regression	100.669689
Lasso Regression	100.477016
Ridge Regression	100.586508
ElasticNet Regression	100.477016
Random Forest Regression	47.424563

3 Classification Model

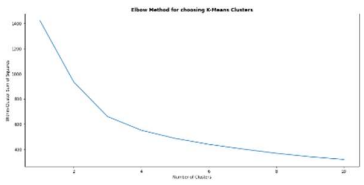
For the classification model, *user_traffic* was created as the target variable from *total_users*. While *total_users* allows bike providers to better predict and meet demand, *user_traffic* aims to help users understand the general level of bike usage so that they can consider alternative commute options based on bike availability. Like the regression dataset, apparent temperature was also dropped from the classification dataset to prevent overrepresentation since it has very high correlation with temperature. Like the regression models, 3 Classification Models, K-Nearest Neighbor and Decision Tree (with and without Pruning), were built and evaluated based on their categorical predictive accuracy. As seen from the table on the left, the Pruned Decision Tree Classification (PDTC) model provides the best predictive accuracy. The PDTC model employs tree pruning using hyperparameter optimization of *max_depth,* which reduces model complexity (thus achieving model parsimony), and *min_samples_split*, which reduces

	Accuracy
K-NN Classification	0.825470
Decision Tree Classification	0.809168
Decision Tree Classification (with Pruning)	0.836210

overfitting in the PDTC model. This was done based on model predictive accuracy. A confusion matrix was also built for each classification model to identify the breakdown of correct classifications. Even though PDTC is the best model (83.6% overall accuracy), it has a significant margin of error for moderate user traffic and is also slightly more prone to overfitting.

4 Clustering Model

For the clustering model, K-Means Clustering was done using temperature, humidity and wind speed. Apparent temperature is used since it relates to users' temperature perception rather than the actual surrounding temperature, which is more relevant for identifying different user segments. Using the Elbow Method as seen on the left, the optimal number of clusters determined was 3. After clustering, Casual and Registered user counts were separately aggregated based on season, hour, holiday, day of the week, working day, and weather type to better profile different clusters. Through the visualizations created in the Jupyter Notebook, there are a few broad trends to be observed. Firstly, Casual users mostly use bikes on non-working days and weekends, compared to Registered users who use it more on working days and weekdays. Casual users also tend to use bikes more in the daytime, compared to Registered users, whose usage peak during mornings and evenings. There is also extremely low usage (for both types of users) across weather conditions 3 and 4 (likely to be extreme weather conditions), and for holidays. For specific clusters, Cluster 1 users mainly use the bike sharing services during Summer and Autumn (determined from dates and temperatures of the corresponding seasons), and during evening hours, whereas Cluster 2 users use the service more during Summer, Autumn and Winter, and during morning hours. For Cluster 3, users are more active during Winter and Spring, and during the afternoon.



5 Further Analysis & Conclusion

The above analysis provides a proper framework for bike sharing service providers to better predict and match service demand through redirecting idle bike resources. This helps to reduce lost customers from bike unavailability, hence improving profit. Bike sharing businesses can also use trends identified from user segmentation to improve service offerings. To encourage usage, discounts can be provided for new Casual users during daytime, on non-working days and weekends. For Registered users (likely working adults or schooling kids), discounts can be given during mornings/evenings, on weekdays, and working days. Appropriate discounts can also be applied for new customers in the respective clusters based on the seasons and time periods these clusters are most active. Through classification, users may also gain insights on bike traffic and availability, enabling them to make better commute decisions.

Despite the given models' ability to predict demand and traffic, there is also a chance it might not generalize well due to overfitting and its inability to account for external demand shocks such as the current COVID-19 situation.