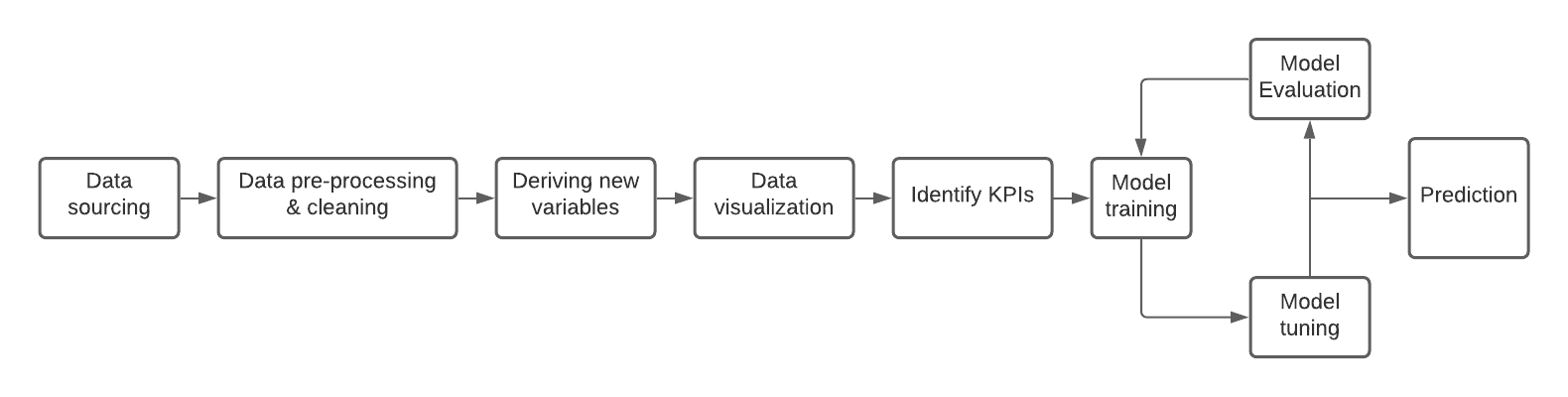
**Memorandum**

Thank you for giving me the opportunity to work on the *EZ Car Rental Company* case. The aim of the project was to predict the rental cost of car using the provided dataset that mainly focused on the features such as journey start time, journey end time, city, and car utilization. This document will describe the methodologies that were applied to fulfil the requirement of the price prediction.

Below diagram briefs the approach.



1. Data Sourcing: The dataset journeys.csv and utilization.csv were read.
2. Data pre-processing:
3. Modified the data types for datetime columns.
4. As the given price were all in dollars, hence, removed ‘$’ sign from the price records.
5. Created a new dataset which comprise of the aggregated available and utilized time in hours for each trip.
6. Dropped records with outlier prices.
7. Derivation of new variables: Derived variables namely TripDurationInHr (Difference between the start and the end journey time), preBookingDurationInHr (Difference between the trip booking and journey start time), isWeekend, yearQuarter, peakHours.

Note: journey creation time is considered as the time when the car was booked.

1. Data visualization: Plotted each variable against the number of trips and the price to understand the demand of rental cars and the associated cost. Also, the different patterns/relations W.R.T. day, hour, quarter was analysed.
2. Identify KPIs:
3. By analysing the graphs, identified the pattern and grouped data into several buckets.
4. Convert categorical variable into dummy/indicator variables.
5. Model training, evaluation, tuning:

As all the features were identified & cleaned, model training was the next step. For this, the dataset was divided into training and testing (20%) sets. As per the data, the case was identified as the regression problem.

Target: price (priceInDollar)

Predictors: tripDurationInHour, preBookingDurationInHour, utilizationInHour, availabilityInHour, isWeekend, yearQuarter1, yearQuarter2, yearQuarter3, yearQuarter4, eveningNonPeakHours, eveningPeakHours, morningNonPeakHours, morningPeakHours, nightHours.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sr. No. | Model | Training Set | | Testing Set | |
| RMSE | R square | RMSE | R square |
| 1 | LinearRegression() | 13.25 | 0.71 | 13.71 | 0.69 |
| 2 | DecisionTreeRegressor(max\_depth = 2) | 13.79 | 0.68 | 13.95 | 0.68 |
| 3 | RandomForestRegressor() | 6.93 | 0.92 | 11.29 | 0.79 |
| 4 | GradientBoostingRegressor() | 10.36 | 0.82 | 10.55 | 0.82 |
| 5 | GradientBoostingRegressor(max\_depth=97, n\_estimators=19, random\_state=0) |  |  | 12.79 | 0.73 |
| 6 | RandomForestRegressor(max\_depth=9, n\_estimators=15, random\_state=0) |  |  | 10.47 | 0.82 |

For hyper parameter tuning ‘Random Search CV was used’.

Model # 6 was selected as the final model as it RMSE is low and R square is high which mean high accuracy.

The model obeys the rules such as increasing the predicted cost with the increase in journey time and utilization. It also predicts for the different cities and peak times. The accuracy score for this model is 82%. The model could be further optimized by tuning the hyperparameters and implementing other cross validation techniques.