PROPOSED SOLUTION

Data Preparation

From the timestamp column in pings.csv, I extracted year, month, day, hour, and minute columns. Then I prepared the target column(online_hours) based on these columns. Then I joined pings.csv with customers.csv on id to get the customer details columns. Dropped timestamp and id. Dropped the year and month columns since they contained only one unique value. This is my train data.

Similarly, I prepared the test data as well. It already had the target column.

Final columns: gender, age, number_of_kids, day, online_hours

Next Steps

I have leveraged Automated Machine Learning (AutoML) technique to further solve this problem. AutoML helps in automating some critical components of the ML pipeline like feature selection/dimensionality reduction, feature engineering, model development, and hyperparameter tuning.

I have used 3 powerful open-source AutoML libraries:

1. EvalML
2. AutoViML
3. TPOT

Several ML pipelines (formed by taking a combination of transformers and models) were tried on the training data with the help of these AutoML libraries.

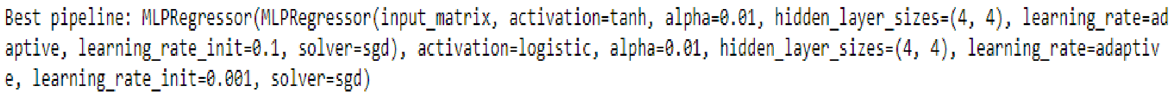
Transformers applied/tried on train data

- *Feature Encoding* - One Hot Encoder technique
- *Missing Value Imputation* - Most Frequent Imputer technique
- *Feature Scaling* - Standard Scaler, MinMax Scaler
- *Feature Selection* - SelectKBest + MIS (Mutual Information Score)

Models applied/tried on train data

- *Gradient Descent Based Models* - Ridge, Lasso, Elastic-Net, SGD, MLP Regressor (ANN)
- *Tree-Based Models* - Decision Tree Regressor, Random Forest, Extra Trees, Gradient Boosting (XGBoost, LightGBM, CatBoost), AdaBoost
- *Distance-Based Models* - KNN Regressor, Linear SVM Regressor

Best pipeline



```
Best pipeline: MLPRegressor(MLPRegressor(input_matrix, activation=tanh, alpha=0.01, hidden_layer_sizes=(4, 4), learning_rate=ad
aptive, learning_rate_init=0.1, solver=sgd), activation=logistic, alpha=0.01, hidden_layer_sizes=(4, 4), learning_rate=adaptiv
e, learning_rate_init=0.001, solver=sgd)
```

RMSE on test data

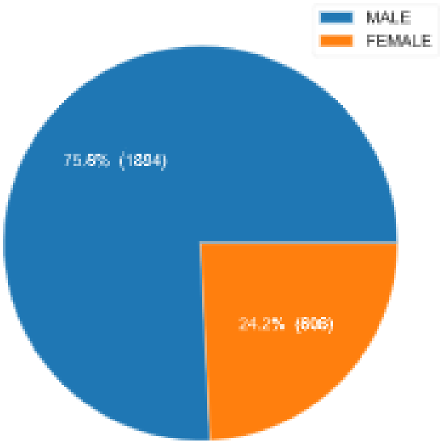
3.102

Note: Error is high because each data point is contributing to some error. Data points in the target column of the test set are discrete values whereas of train set are continuous values.

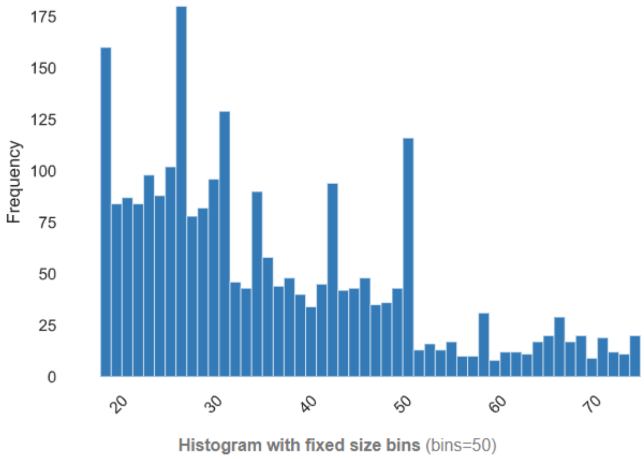
VISUALIZATIONS

Frequency Plots

Gender

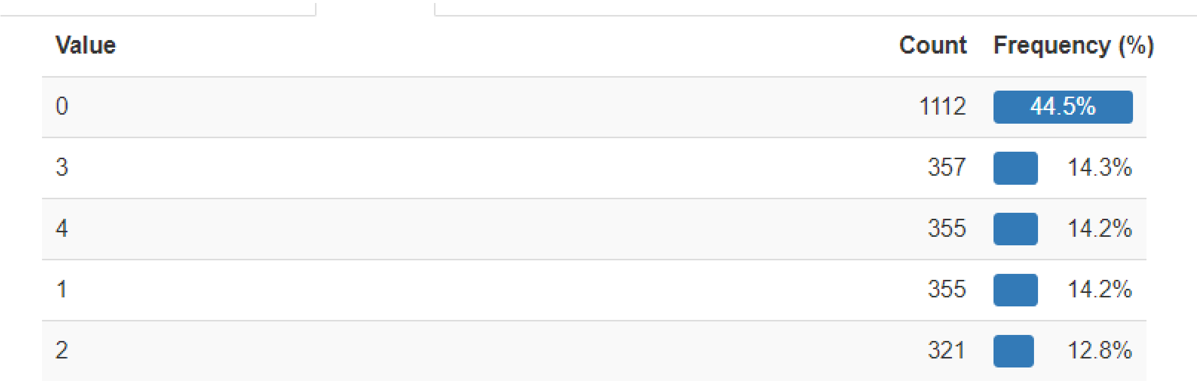


Age



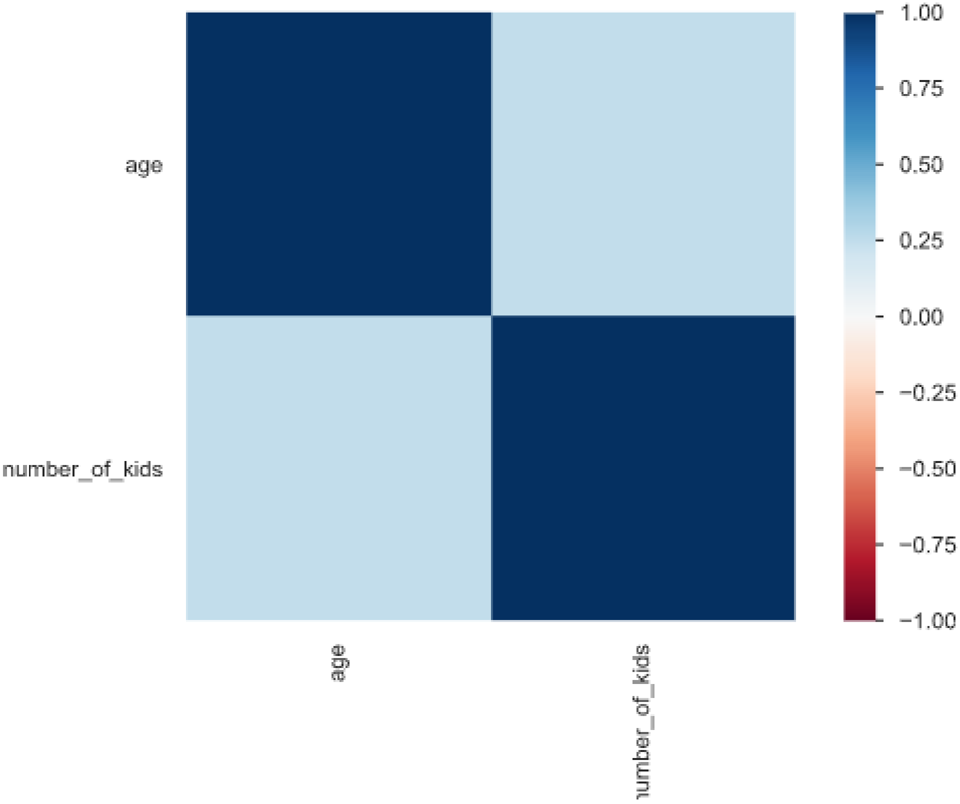
Histogram with fixed size bins (bins=50)

Number of Kids

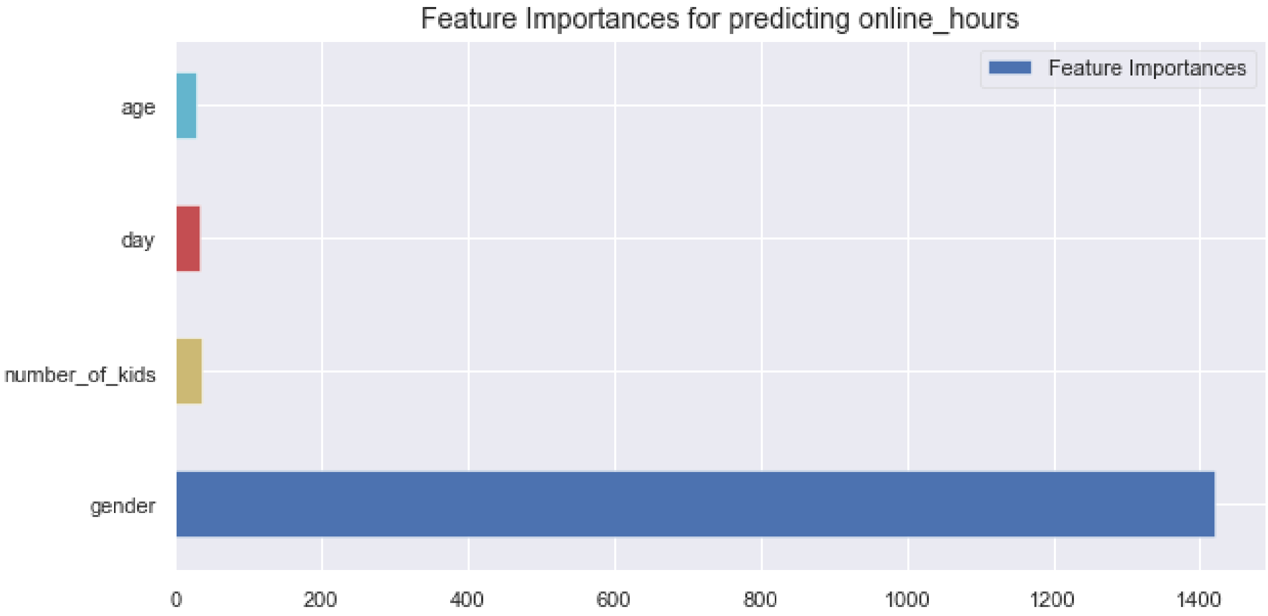


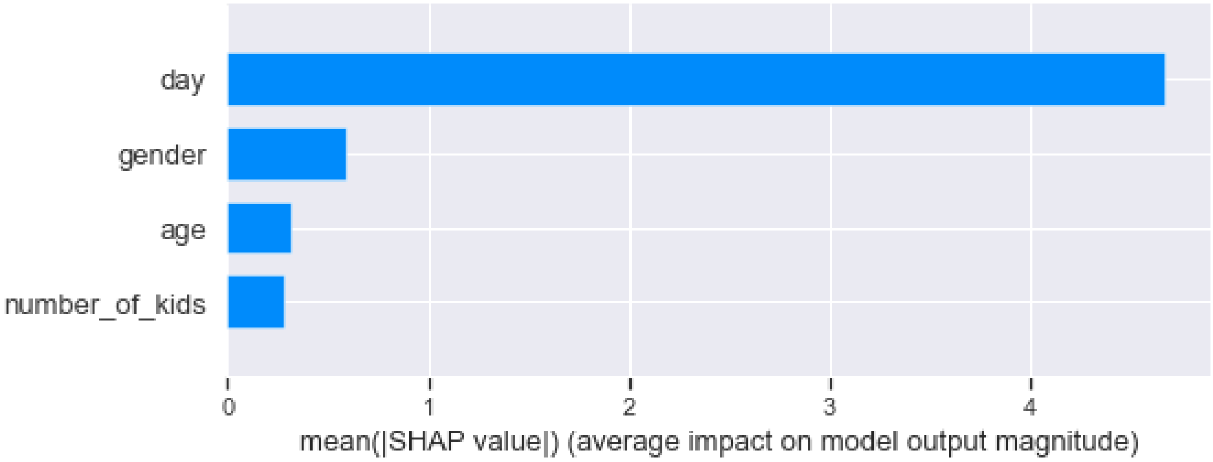
Value	Count	Frequency (%)
0	1112	44.5%
3	357	14.3%
4	355	14.2%
1	355	14.2%
2	321	12.8%

Correlation Plot

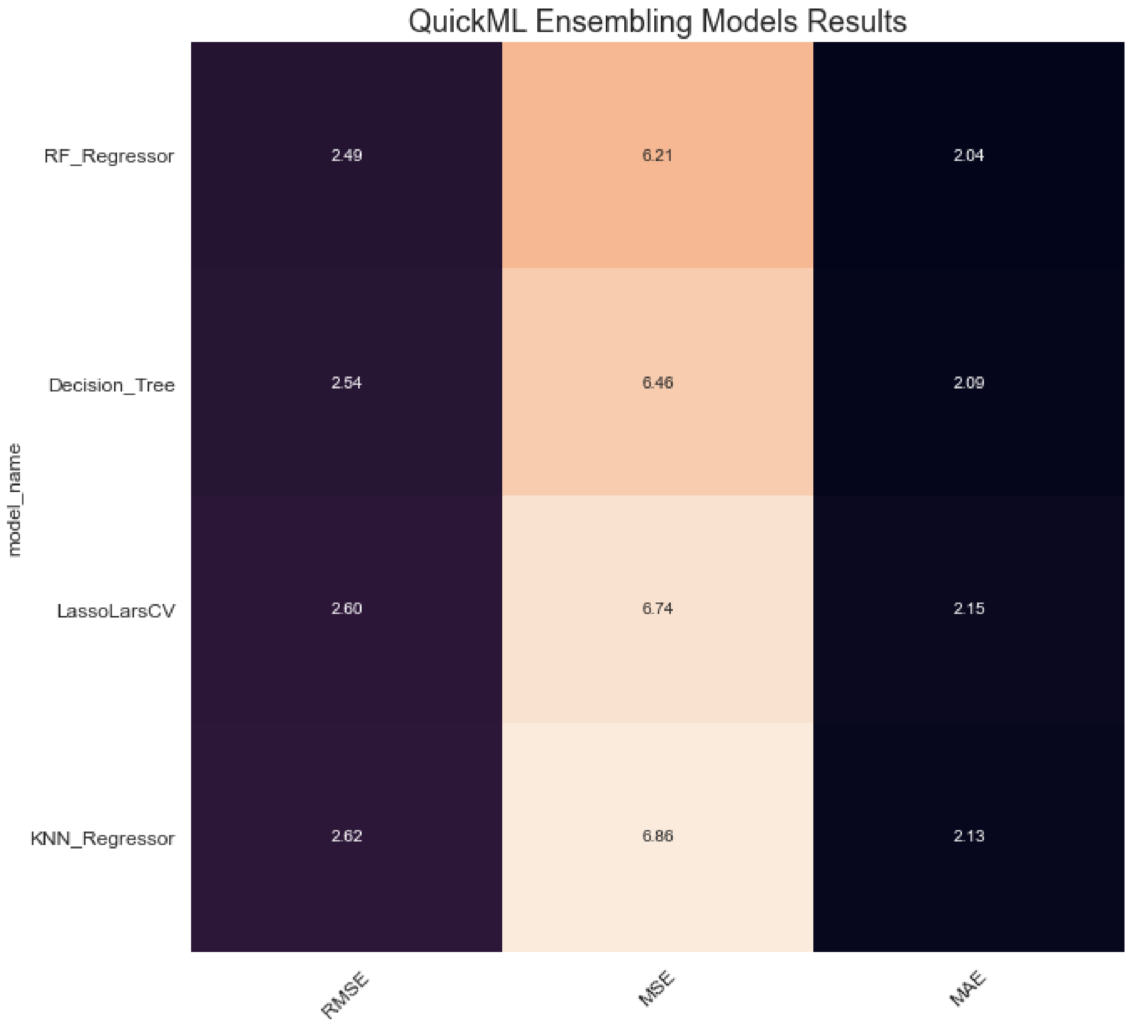


Feature Importance Plots





Result Plots



INTERESTING INSIGHTS

- 76% of the customers of the InvestorAI platform are male.
- Most of the customers of the InvestorAI platform are in the age group of 20-30 (young) followed by 30-50 (middle-aged). There are a few customers in the age group of 50 and above (elderly).
- Approximately 45% of users have no kids.
- 1-month training data is given i.e., of June 2017.
- The gender column has the highest feature importance.

REFERENCES

https://evalml.alteryx.com/en/stable/user_guide/components.html

https://github.com/AutoViML/Auto_ViML

https://github.com/EpistasisLab/tpot/blob/master/tpot/config/regressor.py