The given problem statement of **predicting customer purchase inclination**, when analysed with the supporting dataset of **1.5 lakh records**, has revealed quite a few **interesting (6) insights** to act upon.

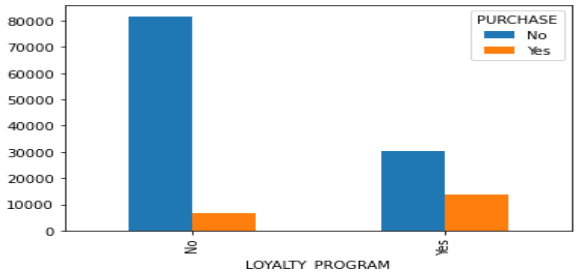
Key Highlights/Assumptions from Classification Methodology:

.

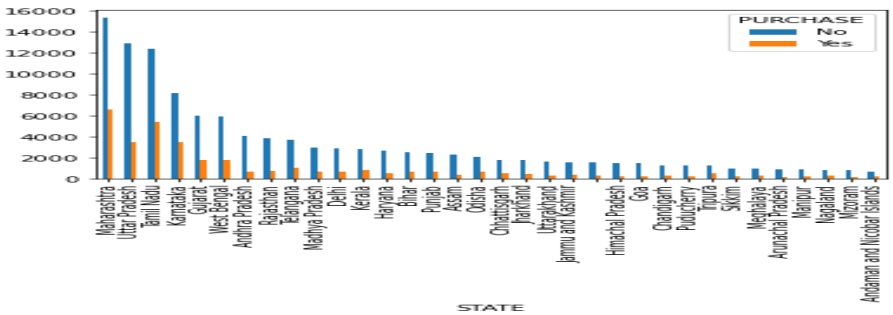
- The data was examined for relevant features and **customer ID** was eliminated since it did not influence customer behaviour and therefore **would not contribute** to model building
- Approximately 12% null values were found in **CUSTOMER_LOYALTY** feature. Since the percentage of null values was less, they were imputed instead of being eliminated. As we do not know whether these 12% customers had access to loyalty program, these values were imputed with **'N.A'**
- **Exploratory Data analysis** was performed on each of the 7 features and relevant insights were drawn from each
- The column contained two continuous numeric features- **AGE and PAST_PURCHASE**. Both of these features were checked for outliers. Although did had outliers, all these outliers were of valid values since there were customers who bought expensive items once a while
- The remaining features were categorical and hence they were label encoded with numeric values for the model to train
- 20% data was kept separated for model validation and rest 80% was built for training
- The algorithms used were Decision Tree, Random Forest, KNN, Adaboost, Gradient Boost and Support Vector Machine. Each algorithm was fine-tuned and cross validated for hyper parameters and best parameters of each algorithm was selected.

Insights:

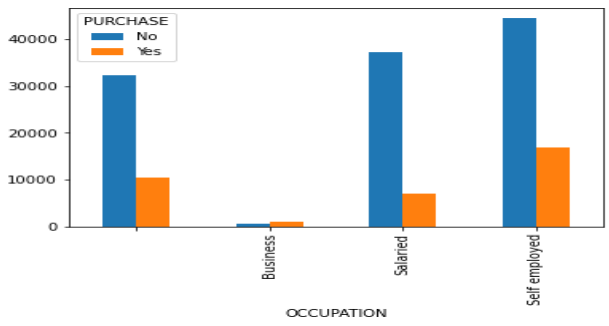
1. The people who **have loyalty program have purchased more** no. of times than people who do not have loyalty program



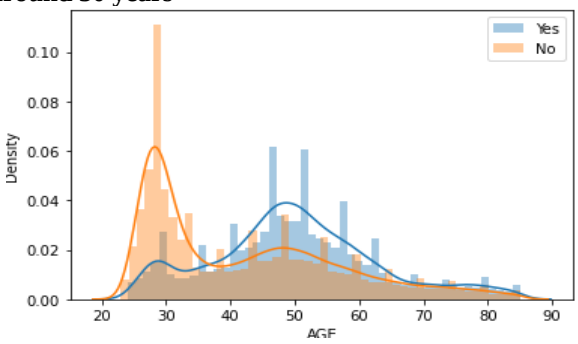
2. Maximum purchase has been made **from 4 states**- Maharashtra, Uttar Pradesh, Tamil Nadu and Karnataka



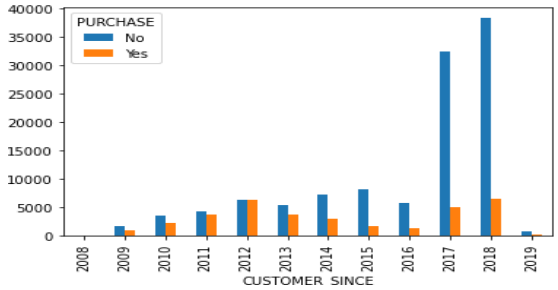
3. Maximum customers are from **Self-employed category,** who have also made maximum purchases. Business category people have made the least purchase



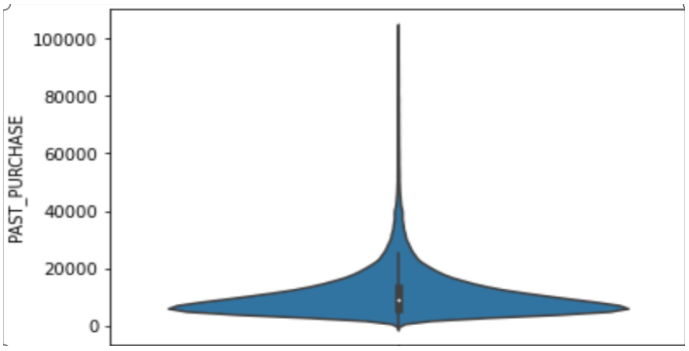
4. **Maximum** purchases have been made in the age group **around 50 years**. **Minimum** purchase has been made in age group **around 30 years**



5. Maximum purchases have been made from people who **have been customers since 2012**



6. 50% of the purchases have been made for the products priced between INR6K-13500



Following were the cross validation scores:
Decision Tree cv score 0.852
Random Forest cv score 0.8571
KNN cv score 0.8339
SVM cv score 0.79285
Gradient Boost cv score 0.8581
Adaboost cv score 0.8548

CONCLUSION:

- All the six fine tuned models were compared based on metrics which included Accuracy, f1_score, sensitivity, specificity and ROC area.
- Based on above metrics Random Forest was selected for further processing.
- This model was then used on the 20% validation data that was kept separated.
- This model gave an f1_score of 86% on validation which was almost equal to the cross validation score.
- This further validated that the model was not overfitting or under fitting.
- This model can predict with 86% accuracy if a customer is most likely to make a purchase or not.
- Finally the model was deployed on test data to create csv file for model accuracy.