Advertisement Success Prediction

Binary Classification Problem

Prepared By: Hiren Kelaiya

Index

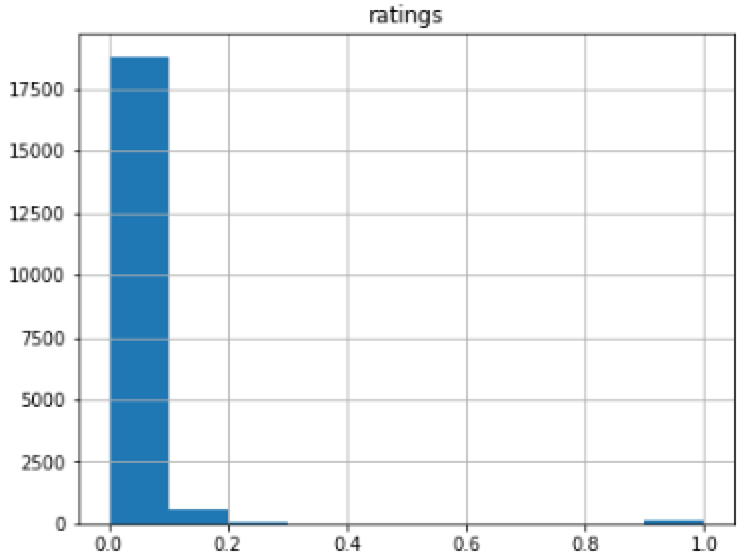
- Initial Data Observation
- Exploratory Data Analysis (EDA)
- Feature Encoding
- Class imbalance handling and Train/Val split
- Machine Learning Models
- Evaluation metric
- Conclusion

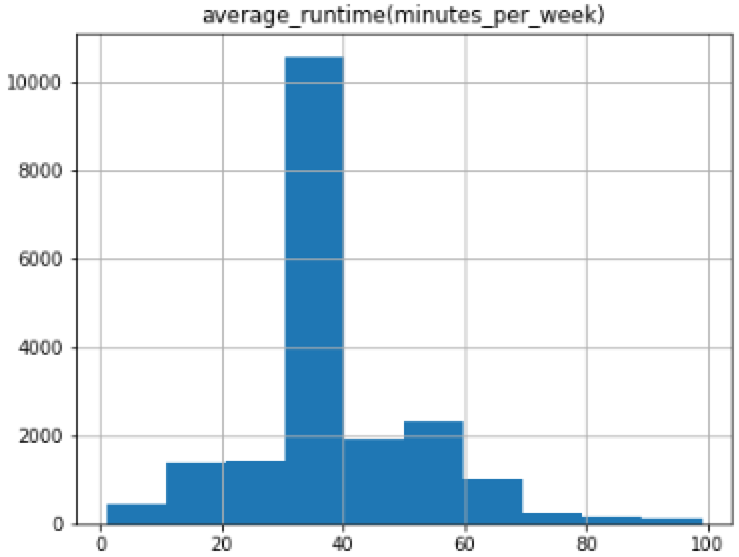
Initial Data Observation

- Training data has 19536 entries while testing data has 6512 entries
- Dataset contains 11 features and 1 target variable. Out of them, only 3 features are numeric while others are categorical.
- None of the features have missing values
- Target variable has imbalanced class distribution
- Scaling is not required as all the numerical features are in similar scale

Exploratory Data Analysis (EDA)

- Univariate analysis of numerical features

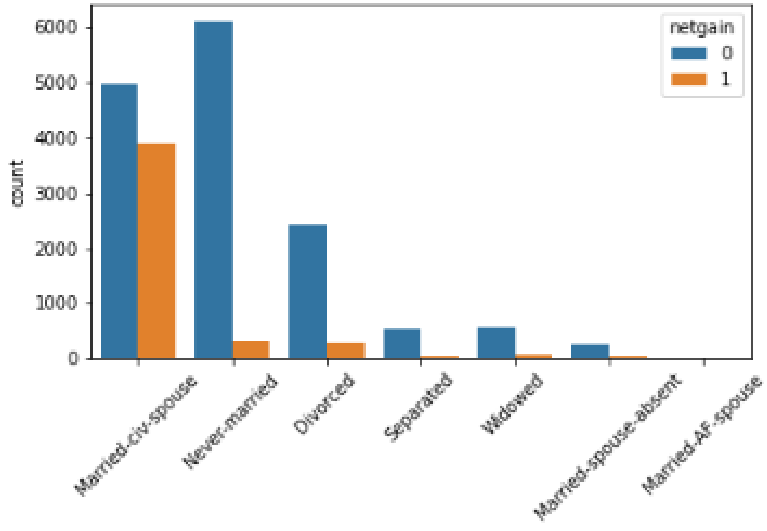


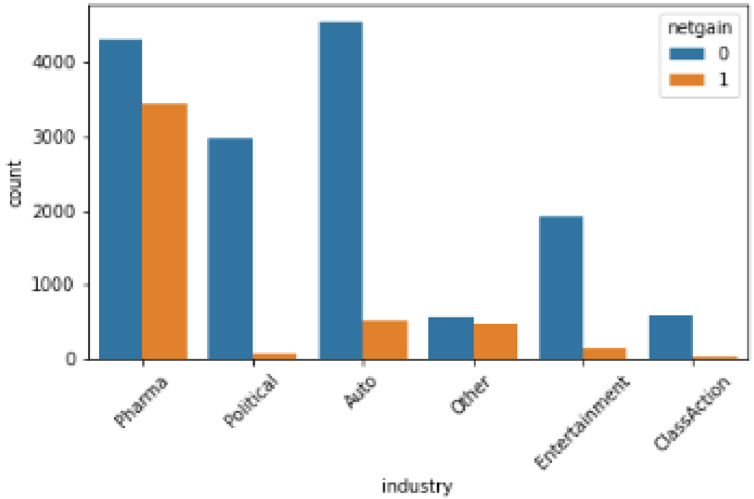


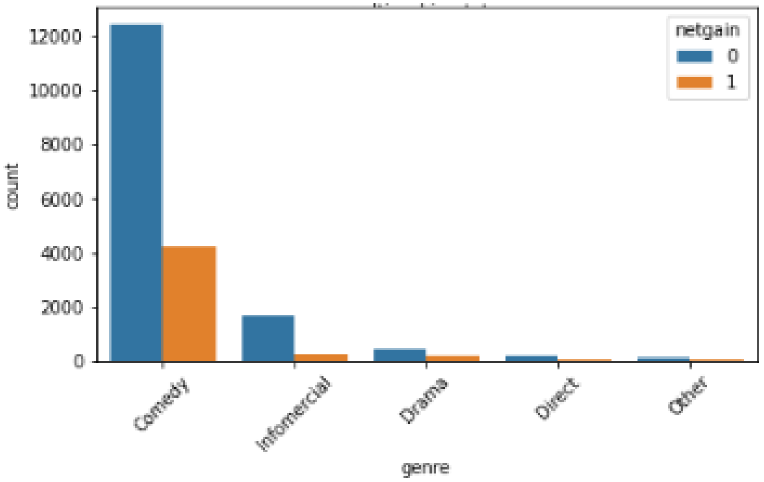
1. The distribution of **ratings** is heavily right-skewed distribution. However, if we look at the quartiles, we clearly see the **25th, 50th and 75th** percentile have the same values. Further observation says that **0.027465** have most repetitive values in the data, hence transformation of this feature doesn't help much. We will keep this as it is.
2. The distribution of **average_runtime(minutes_per_week)** is more or less normally distributed.

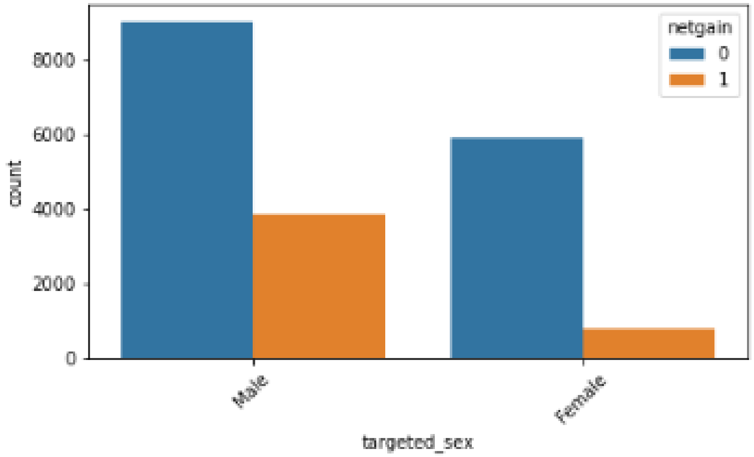
Exploratory Data Analysis (EDA)

- Bivariate analysis of categorical features



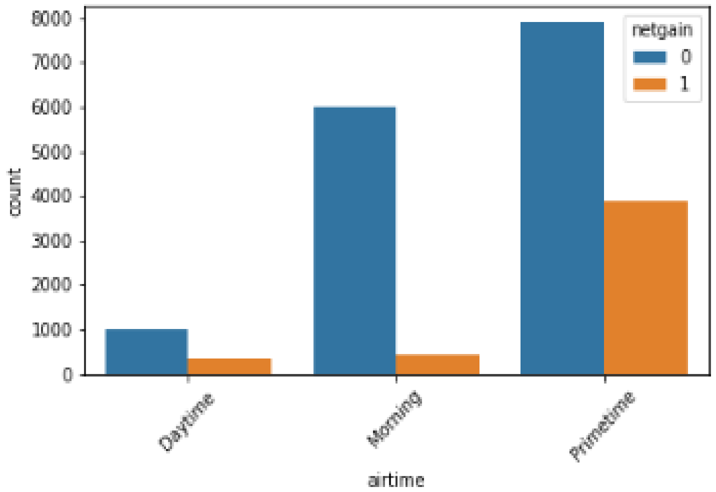


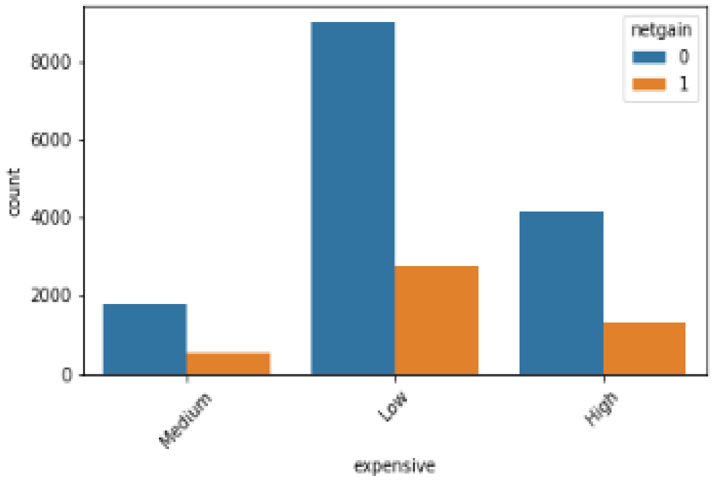


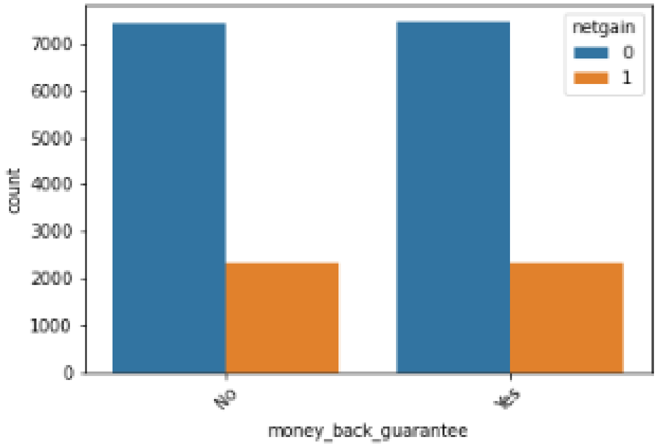


Exploratory Data Analysis (EDA)

- Bivariate analysis of categorical features







1. **Relationship status:** It is clear from the above bar graph that married audience lead to higher net gain compare to others.
2. **Industry:** It seems pharma industry has highest net gain when ad is sold.
3. **Genre:** Comedy genre leads to higher net gain when ad is sold compare to other genre.
4. **Targeted sex:** It seems that if the targeted sex is male, it will lead to higher gain compare to female.
5. **Air Time:** It is clear from the bar graph that primetime will lead to highest gain when ad is sold.
6. **Expensive:** Less expensive ads will lead to highest gain when it is sold.
7. **Money back guarantee:** It seems money back guarantee doesn't lead to overall net gain when ad is sold.
8. **Air Location:** It seems US is the potential country that leads to higher net gain when ad is sold.

Feature Encoding

- Here, following techniques are used to encode the categorical features:

1. **Label Encoding** used to encode ordinal feature ***expensive*** and two nominal features ***targeted_sex*** and ***money_back_guarantee***.
2. **Dummy Encoding** used to encode nominal feature like ***airtime, realtionship_status, genre, industry***.
3. **Binary Encoding** used to encode few nominal features such as ***airlocation***. The reason why it is used because of these features have multiple categories. One-hot encoding or dummy encoding is not an optimum choice when we have multiple categorical features having multiple categories. The binary encoding works really well in case of a high number of categories.

Class imbalance handling and Train/Val split

- The target variable ***netgain*** has imbalance between two classes.
- Class **1** needs to be up-sampled in order to make it equivalent to the class **0**. **SMOTE** technique is used to handle class imbalance.
- Here, **80:20** splitting of training and validation data lead to higher ***f1 score*** compare to other splits.

Machine Learning Models

- Tried following few classifiers and selected the best performing model for further tuning and testing.
  1. Logistic Regression
  2. Decision Tree
  3. Random Forest
  4. XGBoost
  5. Support Vector Classifier
  6. Linear Support Vector Classifier
  7. K-nearest Neighbors
  8. Stochastic Gradient Descent (SGD) Classifier

- Out of these models, **XGBoost Classifier** outperformed all others with having **F1 score** of **62.0** on validation data.

Evaluation metric

- For this problem, **F1 score** is used to evaluate the model on test data.
- F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it’s better to look at both Precision and Recall.

- F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Conclusion

- This is the binary classification problem with imbalanced target classes.
- All the categorical and numerical features are correlated to the target variable ***netgain*** except **money back guarantee** which does not have much of an impact. Also **airlocation** is removed due to unimportance in predicting target variable.
- **XGBoost Classifier** performs better after tuning compare to other known classifier with **F1 Score** of **62.0** on validation data.

Thank you