Retail Marketing Take Home Assessment

1. Overview

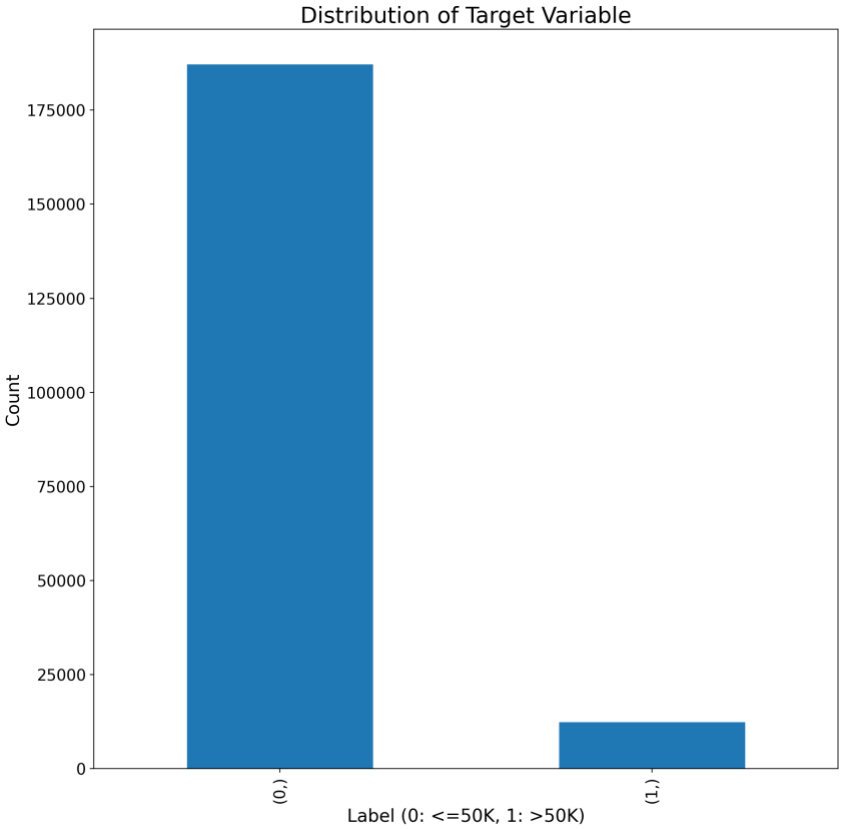
The project comprises of 2 parts

1. **Predictive Modelling:** The goal of this predictive model is to predict if the salary of the given person exceeds $50K or not
2. **Segmentation Analysis:** The goal of the segmentation analysis is to identify groups using the US census data for marketing purposes.

2. Data

This data set contains weighted census data extracted from the 1994 and 1995 Current Population Surveys conducted by the U.S. Census Bureau. Each line of the data set (censusbureau.data) contains 40 demographic and employment related variables as well as a weight for the observation and a label for each observation, which indicates whether a particular population component had an income that is greater than or less than $50k.

This dataset contains 42 columns (including target) and 199,523 rows of data. There are 187,141 people who earn less than $50K and 12,382 people who earn over $50K.



3. Predictive Model:

The goal of the predictive model is to effectively predict if a person earns less than or more than 50K given the US census data. The dataset consisted of 28 categorical features and 13 numerical features. Logistic Regression, Random Forest, and CatBoost models were utilized in several experiments to determine the best model.

Numerical Data Processing

Correlation between different numerical features was calculated. If two features were highly correlated (Pearson correlation over 0.7), one of them was dropped. This was done to prevent duplicating signal that could skew the model results. The columns dropped in this step were ***detailed industry recode*** and ***num persons worked for employer***.

Categorical Data Processing

For categorical features, there were two steps conducted.

1. **Missing Data Imputation:** Missing data was found in the column ***Hispanic origin*** this was imputed by the value 'Do not know'
2. **Target Encoding:** Target encoding was conducted to replace categorical features with numeric value which keeping a sense relation to the target variable. This was done for models that do not inherently support categorical variables.

Imbalanced Data

Since we know that the data was imbalanced, we ran some experiments with synthetically generated data using SMOTE. This technique draws a line between two data points from the minority class and randomly picks a point on that line to generate a new data point. This process is repeated until the number of data points for minority class is the same as that of the majority class.

Data Scaling

For Logistic Regression model, the data was once again scaled between -1 to 1 after all categorical features were replaced by their corresponding numeric values. This was done so that the coeffcients (feature importance) for the model is interpretable and is not skewed by the magnitude of the values

Experiments

Several modelling experiments were run with different conditions and their results have been noted. It should be noted that the first 3 models were used to determine the optimal parameters using **Hyperparameter Tuning**. In an ideal situation, these parameters would be calculated for each model but due to time and computation limitations the process was not conducted.

Model	valid_loss	test_loss	valid_f1_score	test_f1_score
Logistic Regression	0.129000	33.153302	0.116889	0.116845
Random Forest	0.117000	0.117200	0.525166	0.529992
CatBoost	0.113000	0.112668	0.557604	0.568643
Logistic Regression (SMOTE)	0.284000	0.285631	0.446480	0.451385
Random Forest (SMOTE)	0.134000	0.134641	0.543698	0.543937
CatBoost (SMOTE)	0.117000	0.116027	0.573012	0.574416
Ensemble (10 Random Forest models)	0.119145	0.116191	0.528185	0.533865
Ensemble (10 CatBoost models)	0.114422	0.111325	0.562801	0.577346
Ensemble Combined (above combined)	0.116784	0.378646	0.545493	0.620178

Based on the results, my recommendation is to proceed with the final model, ensemble of 10 Random Forest + 10 CatBoost models. This might not be the most optimal solution in the development steps however, it is the most robust solution due to the combination of a 20 'weak' models.

4. Segmentation Model:

The goal of the segmentation model is to create groups that can be targeted for marketing purposes using the US census data. Since the results will be used for marketing purposes, it is important to make sure that our clusters are interpretable and can be explained to non-technical audiences.

Categorical Feature Processing

The categorical features are processed in two steps

1. **Ordinal Encoding:** This is done for two variables, ***education*** and the target (henceforth referred as ***salary***) variable. These two variables were chosen because there is an inherent order to the categorical values present in the data.
2. **One Hot Encoding:** This was done for the rest of the data since we are trying to cluster similar groups together. One Hot Encoding would create new columns with values equal to 0 or 1 for each category in the data. This is especially helpful

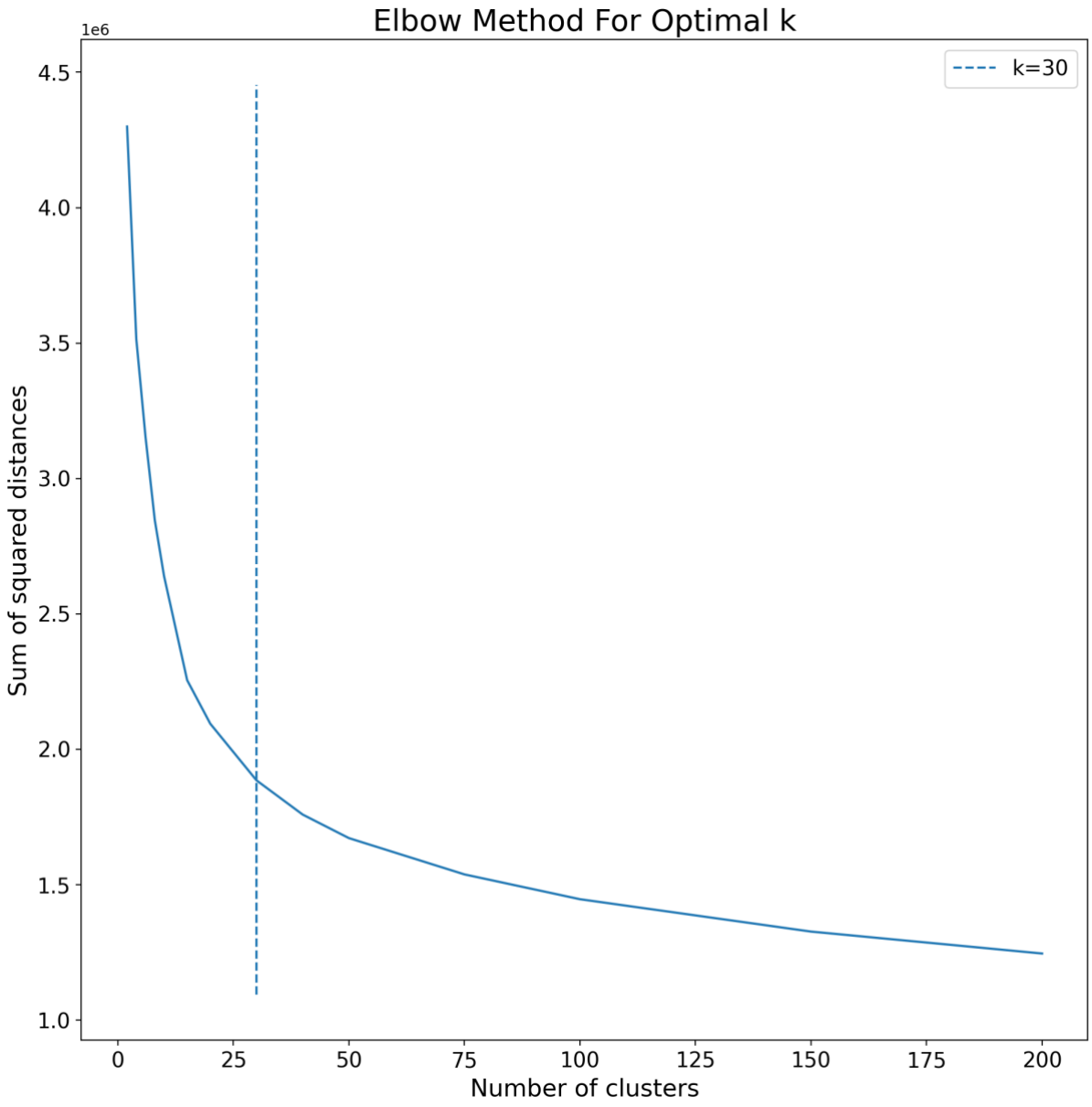
for our case because it maximizes the penalty when the groups are different and helps the model cluster better.

Numerical Feature Processing

Numerical features (including the ordinal encoded features) were scaled to a value between -1 and 1 so that large difference in the magnitude of values does not distort the distance calculations during clustering

K-Means Clustering

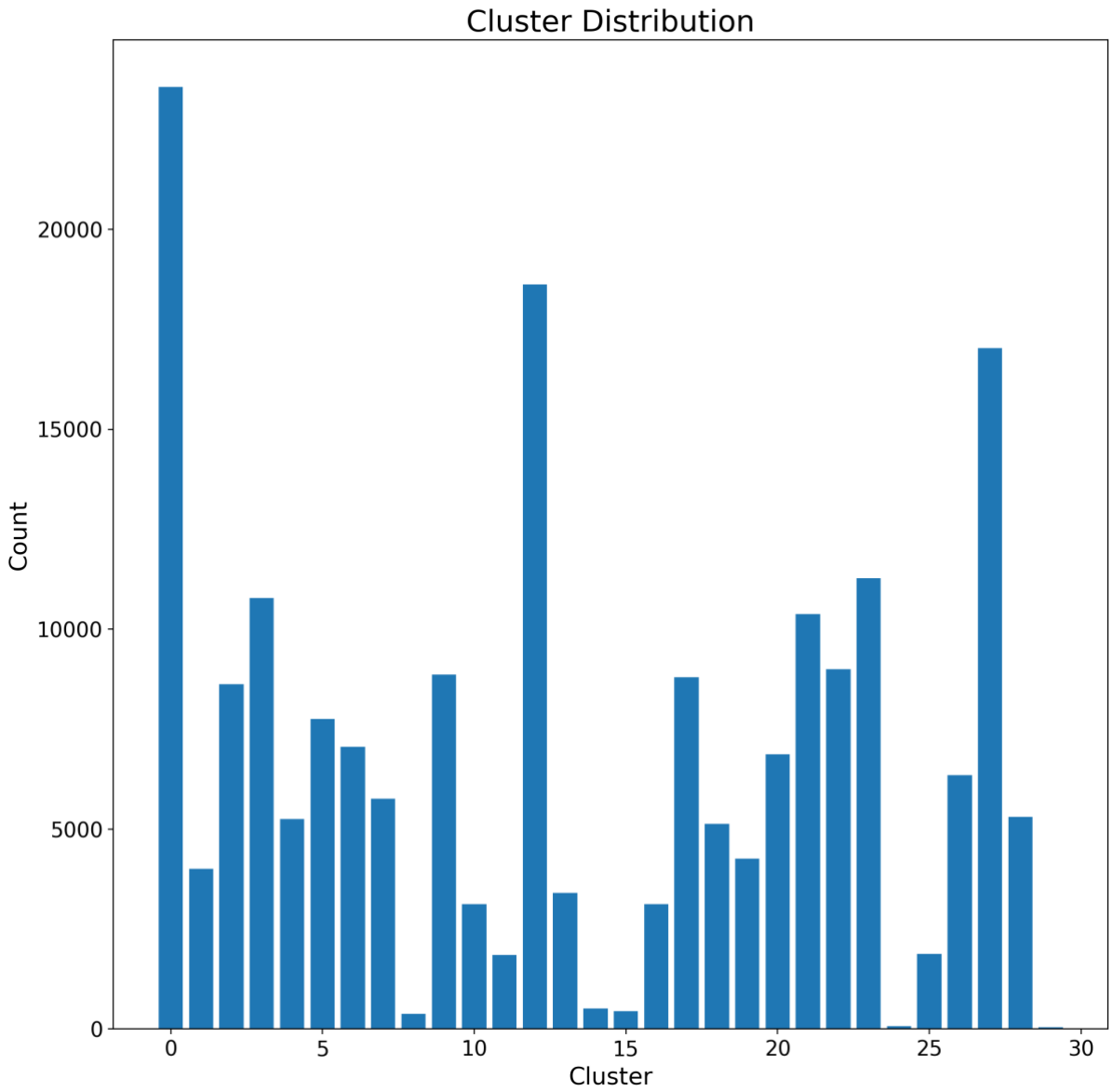
To determine the optimal number of clusters, K-Means clustering was used with 14 different sizes of clusters. The sum of squared errors was calculated for each of these 14 experiments, and it was visualized as shown below.



From the figure, the elbow occurs at 30 clusters which will in turns becomes the optimal number of clusters that we will be using in our segmentation algorithm.

Cluster Analysis

The clusters were created using K-Means where k=30 and the distribution is as follows



These clusters are further categorized into meaningful categories in the ***cluster_profiles.xlsx*** sheet.

Only the summary and the Cluster ID 0, 1, and 2 have been formatted but the raw data for all the clusters is present in the sheet. Sheets only contains general information about demographics such as age, monetary information such as different forms for

income and loss, and work-related information such as number of weeks worked in a year. In addition to the above information, each cluster has additional information regarding its representation (information that applies to atleast 90% of the cluster)

This information is interpretable to non-technical audiences and can be used effectively for marketing purposes.

Example from ***cluster_profiles.xlsx***

Cluster ID	Cluster Size
0	23,554

Demographics

Category	Avg	Range
age	7.03	0-14

Category	Majority Class	% in Majority Class
education	Children	100.0%
country_of_birth_self	United-States	95.4%
citizenship	Native- Born in the United States	95.4%

Monetary

Category	Avg	Range
wage_per_hour	0	0-0
capital_gains	0	0-0
capital_losses	0	0-0
dividends_from_stocks	0	0-0

Category	Majority Class	% in Majority Class
salary	- 50000.	100.0%
tax_filer_stat	Nonfiler	100.0%

Work

Category	Avg	Range
weeks_worked_in_year	0	0-0

Category	Majority Class	% in Majority Class
class_of_worker	Not in universe	100.0%
detailed_industry_recode	0	100.0%
detailed_occupation_recode	0	100.0%
major_industry_code	Not in universe or children	100.0%
major_occupation_code	Not in universe	100.0%
member_of_a_labor_union	Not in universe	100.0%
reason_for_unemployment	Not in universe	100.0%
full_or_part_time_employment_stat	Children or Armed Forces	100.0%
own_business_or_self_employed	0	100.0%

5. References:

- K-Means: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- Hyperparamter Tuning: https://catboost.ai/docs/en/concepts/parameter-tuning
- Optimal no of clusters: https://stackoverflow.com/questions/51762514/find-the-elbow-point-on-an-optimization-curve-with-python
- Scaling: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html