COMP6200 Data Science Semester Project

Andre S

Problem Statement

In this project, we want to find the answer to the following question:

Given data that is not geographical, is it possible to predict whether a location, in this case a suburb is near beach or not?

Data Sources

In this project, we use several tables from Australian Bureau of Statistics 2016 census. Every table's rows consist of suburbs in Sydney. The tables are:

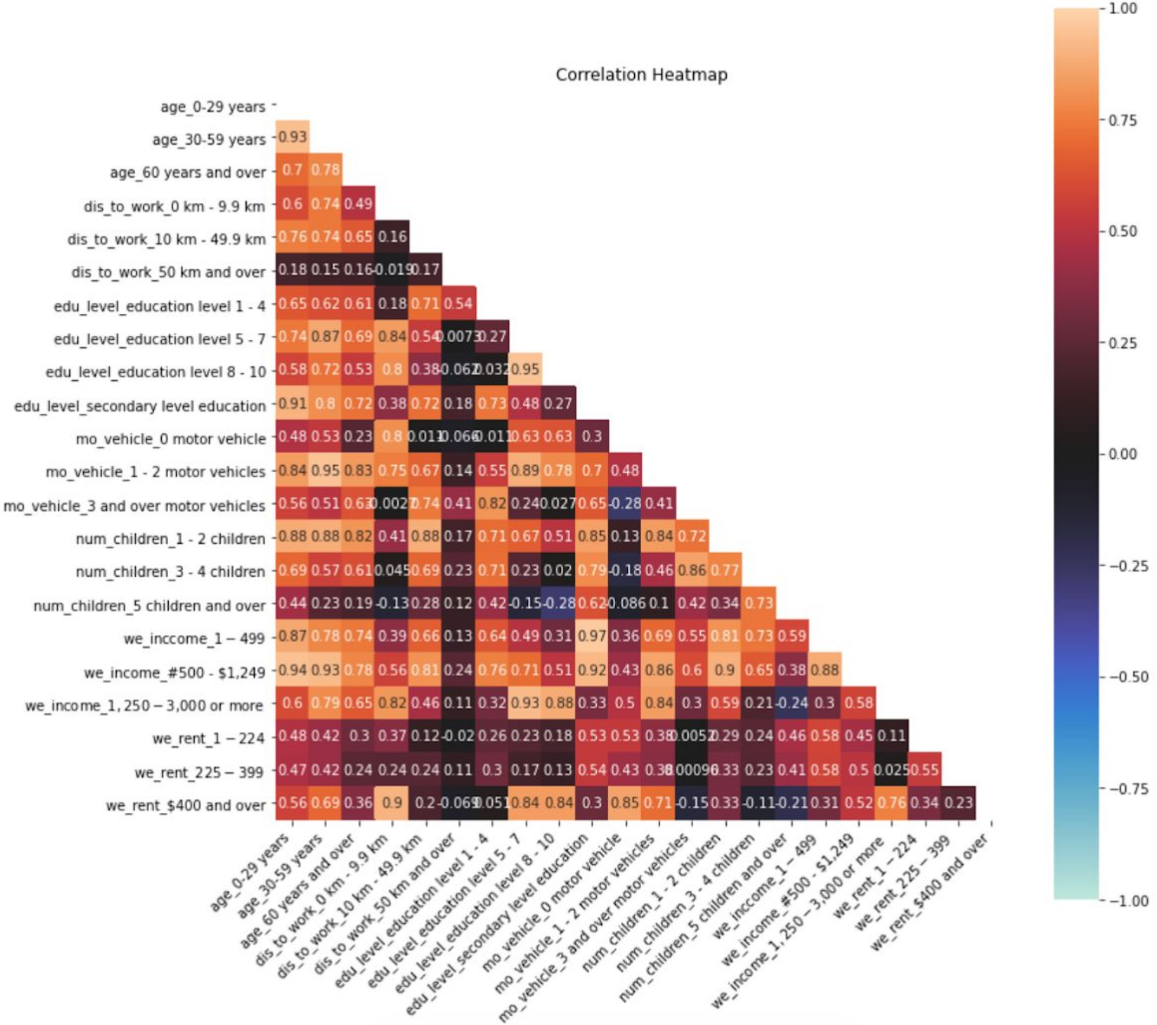
1. Cultural Diversity, counting persons, 10 years age group
2. Cultural Diversity, counting persons, total personal income (weekly)
3. Employment, Income and Education, counting persons, distance to work
4. Employment, Income and Education, counting persons, education level
5. Selected Dwelling Characteristics, counting dwellings, rent (weekly)
6. Selected Dwelling Characteristics, counting dwellings, number of motor vehicles
7. Selected Family Characteristics, counting families, count of children

Method Used

In this project, all of tables are loaded into jupyter notebook and all of them are combined into a single dataframe. Every last row and and last column which contains tables' summary is stripped and the name of the first dataframe is combined_df. For correlation study, features in combined_df are combined using simple summation so the number of columns reduced from 72 in combined_df to 22 in merged_df. Lastly, for classifier tuning, several dataframes that contain optimized features are created to for tuning purpose. The models that are used in this project are Support Vector Machine, Logistic Regression and Random Forest.

Results (1)

One of the highest correlation is number of persons with weekly income 1 - 499 dollars and number of persons with secondary education level, the correlation score is 0.97.



Results (2)

Every model is trained using different kind features:

1. Trained with all available features (72 columns)
2. Trained with combined features (22 columns)
3. Features that are highly correlated with target
4. Features selected using SelectKBest

For Support Vector Machine, the highest cross validation mean score is achieved when the classifier uses all of the available features and the highest cross validation mean score is 0.8475. The highest cross validation score for Logistic Regression is 0.8654 and condition to achieve that is to use dataframe with features which correlation score is over 0.15. For Random Forest classifier, the highest cross validation mean accuracy score is 0.8405 and it is achieved when the dataframe which is used for learning has 7 features selected by SelectKBest.

Conclusion

Given data that is not geographical, is it possible to predict whether a location, in this case a suburb is near beach or not? The answer is “no”