Intelligent Systems Project Report

Fabio Cirelli, University of Insubria
2022 - 2023 Academic Year

Regression models

Regression models - dataset

This part of the project aims at modeling the renting prices of households in India with two different techniques.

The dataset that was used included different features of a household, such as its size, the number of available bathrooms, halls and kitchens, information about the floor, and so on.

Regression models - dataset

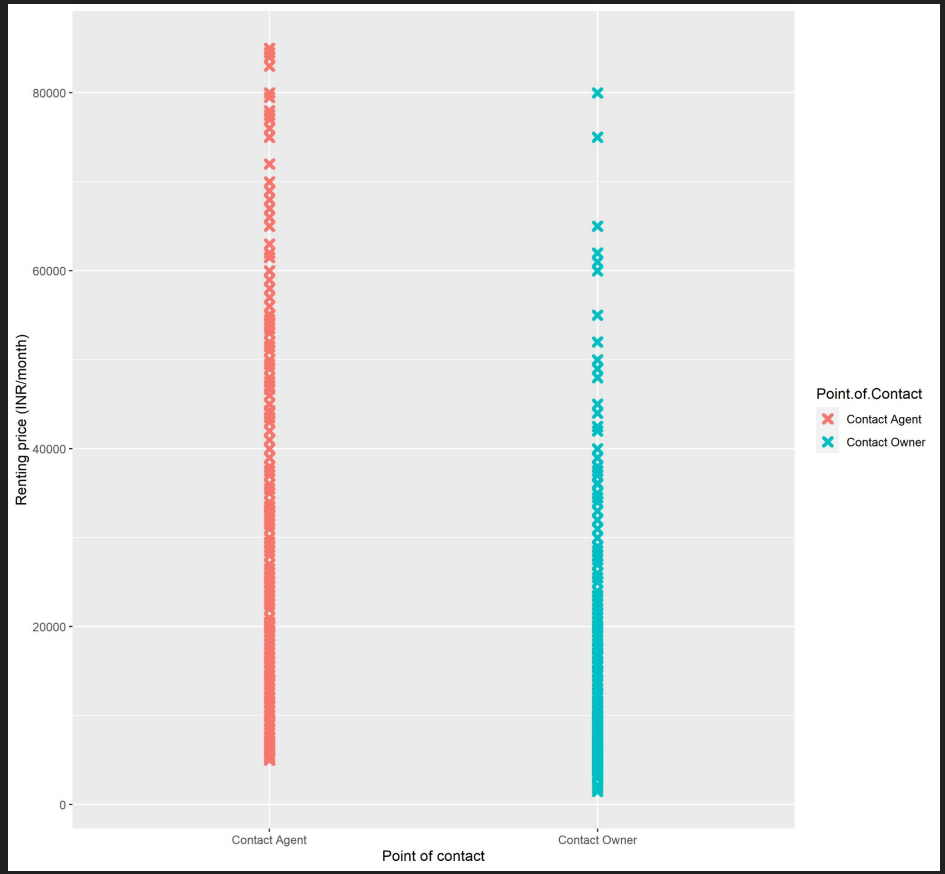
Most of those features turned out to be correlated with the renting price, though, the features with the highest correlations are:

- the number of bathrooms
- who to contact for more information (point of contact)
- the total number of floors
- the size of the household

Some other features, such as the furnishing status, the type of tenant preferred or the area of the city have a very low correlation with the price of a house.

Regression models - dataset

One thing to note about the point of contact is that households for sale by agents tend to have a higher renting price compared to households for sale by private owners



Regression models - feature selection and outlier removal

After manually removing features with a poor correlation, a feature selection procedure was carried out and it removed one more feature: the number of total floors.

Furthermore, data that were too influential (i.e. outliers) were removed in order to obtain more correct models.

Regression models - performance metrics

Finally, the dataset was split: 80% of the data were used as a training set and the remaining 20% were used to evaluate the models' performance, with the following results

	Root Mean Square Error	Mean Absolute Error	R^2
Ordinary Least Squares	7955.7	5705.0	74.6%
Least Median Squares	37367.7	13947.1	38.6%

Classification models

Classification models - dataset

This part of the project aims at predicting whether a patient has diabetes or not with two different techniques.

The dataset that was used only included measures obtained on female patients with an age of at least 21. Only ⅓ of the patients in the dataset had diabetes, but no balancing was carried out in order to not introduce bias.

Classification models - dataset and feature selection

For each observed patient the dataset included the number of pregnancies they had, measurements of the glucose and insuline level, their blood pressure, their age, their health status (i.e. whether the patient has diabetes or not), and more.

The features were selected with the aim of minimising the probability of misclassification. Only three features survived:

- the number of pregnancies
- the glucose level
- the Diabetes Pedigree Function (DPF, likelihood of diabetes based on family history)

Classification models

The dataset was then split in a training set and a test set with an 80 to 20 proportion. The models' results were compared with the actual health status of the patients in the test set, with the following results:

	Accuracy	Precision	Recall	True Negative Rate
Logistic Regression	78.87%	81.55%	88.42%	59.58%
Neural Network	80.28%	83.17%	88.42%	63.83%