Project Report

Starting Position

Client

The client is a fictional Indian government agency that aims at increasing the life expectancy of Indians in the 9 states it is active. The name of the agency is Life Expectancy Action Agency.

Task

As my client wants to improve the average life expectancy of average Indians, it needs a machine learning model that predicts the life expectancy of Indian individuals. The agency wants to allocate its resources depending on the potential life expectancy of individuals. More precisely, the smaller the life expectancy of a person the more money and time should be invested in order to increase the life expectancy of that person. My task is to create such a model if possible and to give further recommendations. For this task, the agency has performed a survey and provided me with the resulting dataset.

Dataset

The survey was conducted in 9 Indian states (the states in which my client is active). The questioned people are members of households in which somebody recently died. They were asked about this deceased person. Each observation (row) represents one deceased person and information about him and his household. It is important to note that this dataset describes the deceased person exactly at the time they deceased. Initially, the dataset contained about 770'000 rows and exactly 121 columns.

Project Process

Initial findings and decisions about variables

Before I started the cleaning process I took an initial look at the dataset in order to get a first impression of the data. I used the data dictionary data dictionary associated with the dataset and some web research to find out what the variables mean. There are 121 variables which is too much to describe all of them here. They can, however, be thematically divided into five different groups:

1. Variables that can be used in order to create the target variable (The age each individual person died)

2. Variables which are about the death of each deceased and it’s circumstances, for example where the death occurred

3. Variables which are about each deceased himself or herself and how he lived, for example, the sex of each individual or whether the individual smoked or not

4. Variables about the household or family of each deceased, for example, the number of rooms the household possesses. Variables which are about the individuals, but almost exclusively determined by the household or family of the individuals, are also in this group

5. Finally, variables which either have no meaning or I wasn't able to find out their meaning, for example, different typed of identification numbers.

At this point, I've already made some important decisions for the following data cleaning process. For obvious reasons variables group 5 can not be used as feature variables for the model. Variables of group 1 will be used to determine the target variable but will have to be deleted thereafter. Variables of group 2 about the death and it's circumstances can also not be used, because information about a persons death is not available yet when predicting a living person’s life expectancy. The only groups left are group 4 about the household of each deceased and group 3 about each deceased.

At the first view, the variables of both of those remaining two groups can both be used to predict the life expectancy of each person. But when taking into account the way the survey was conveyed, I came to the conclusion that variables of group 3 about the individual deceased themselves can't be used. The problem is, that the information conveyed in those variables is about the individual just before he died. But when we try to predict the life expectancy of a person, this person is not lying on the deathbed, but it's just a moment of his or her life. The life expectancy of a person correlates with certain factors (There often is causation). Unfortunately for us, the states of those factor at the time of death are not correlated to life expectancy, but to the age of the deceased. In order to illustrate this, I'll make an example. Let's take the categorical variable about a person's drinking habit. A person that never drinks alcohol and never has most likely will die old. A person that has been an alcoholic for 10 years and is sober now, most likely will not get that old as his liver is already damaged. But if we measure the same variable at the time the person is just dying the results will falsely show the opposite. The reason for that is that a person which is very young, let's say a boy, usually doesn't drink. If he dies as a child he will never have had even a sip of alcohol in his life. The older somebody is, however, the more likely he or she will have been an alcoholic once in his life. In order to use this kind of information for my purpose, the survey would have to be conveyed at an equal age for all participants or at a random age, but not at the time the person died.

Fortunately, we still have group 4 about the household of the deceased. We can use those variables for our purpose. The reason that those variables don't fall victim to the problem described above is the following: In India, households are almost always comprised of at least the nuclear family but usually even the bigger family. This also means that they are mixed in age. This, in turn, means that the variables of this group are not correlated to the age of the deceased and can be used as indicators for life expectancy.

To conclude: We will predict the life expectancy of a person given the variables of his or her household.

Data Cleaning

The first and most important step of cleaning was to ensure that each row contains the dependent variable ‘lifetime' which describes at which age the deceased died. In some cases, there was not enough information to compute ‘lifetime' or the information was obviously faulty, so I had to delete those rows. In other cases, there was not enough information to calculate ‘lifetime' precisely to the day. In those cases, I decided not to delete those rows, but to calculate ‘lifetime' as precisely as possible. I ended up with a few very unrealistic ‘lifetime' values, either below 0 or far too old for a real person. I deleted those rows.

Next, I had to delete all columns that fall into any group of variables besides group 4 about the household of the persons. The reason I had to do this is described in the last part of this report.

Furthermore, I had to deal with missing values. I wanted to delete all missing values because they are harmful to my regression model. The process of deleting missing values involved making bar plots of the proportion of missing values among each variable left and histograms of the distribution of missing values per row. I deleted columns that were mainly consisting of missing values. When the histogram indicated that it wouldn’t lead to too many missing rows, I tried and deleted all rows with missing values.

The values of some categorical variables were almost always the same. As such variables can lead to overfitting I had to delete them. Initially, all categorical variables were encoded with integers. For Exploratory Data Analysis, I find strings describing the categories more useful. Therefore I wrote an algorithm which changed the integers to appropriate strings using the data dictionary. This brought to light that there are some categories which have no documented meaning. I deleted all rows with such values.

After the cleaning process, I was left with 30 feature variables and 715'550 rows. Finally, I divided the dataset into a training and a test set.

Exploratory Data Analysis

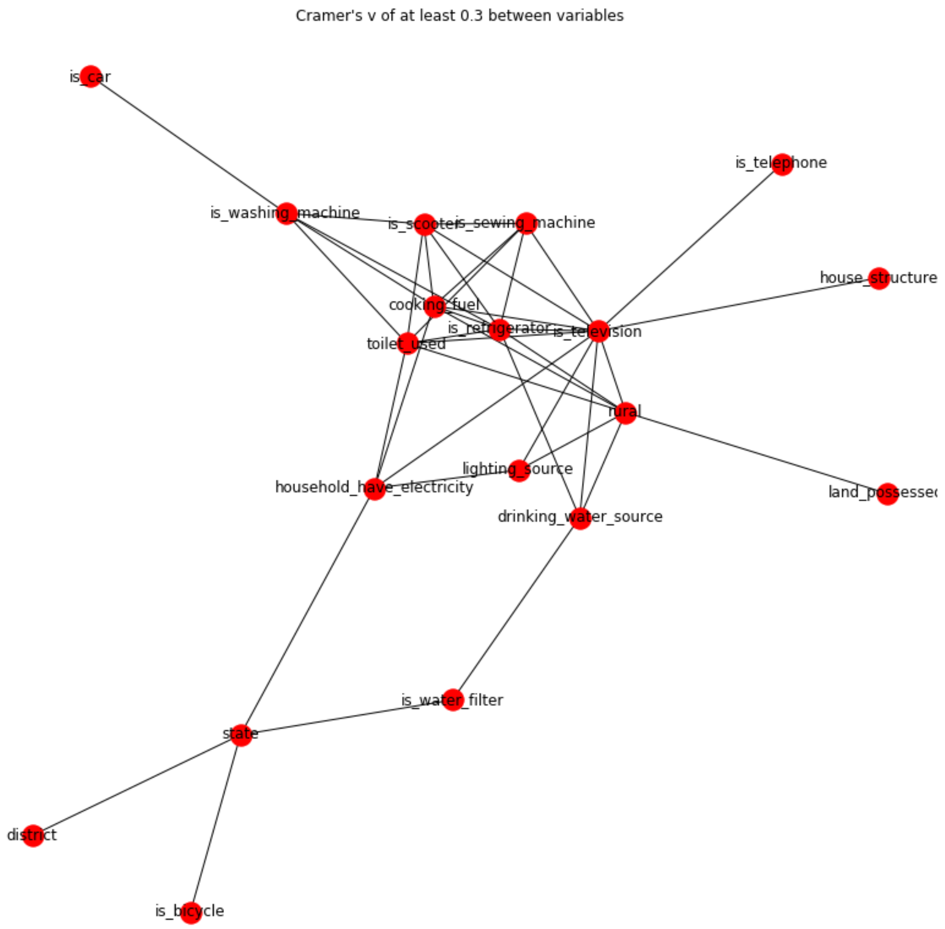
All Exploratory Data Analysis steps I describe in this chapter were done only with the training data. It is important to know for understanding the process that all but one remaining feature variables were categorical.

First I took a look at the distribution of the target variable ‘lifetime'. I expected the distribution not to be Gaussian (but rather right skewed). The histogram and D’Agnostio and Pearson’s normal test confirmed this assumption. Thus I tried to make the distribution normal by taking the natural log, which didn’t work.

Next, I needed to figure out if there are statistically significant differences in life expectancy (‘lifetime') between the categories within each categorical variable. As ‘lifetime' is not normally distributed, a parametric hypothesis test such as ANOVA was not appropriate for this task. I needed to do a non-parametric hypothesis test and chose to perform a permutation test. The Null hypothesis was that there is no statistically significant difference between the mean life expectancy of the category with the highest mean life expectancy and the category with the lowest mean life expectancy. The alternative hypothesis was that there is such a statistically significant difference. As I made one such test for each categorical variable (29 in total), I decided to do a Bonferroni adjustment of the alpha level and used 0.05 divided by 29. The permutation test rejected the null hypothesis in case of all variables. Furthermore, I created a list of the highest difference in life expectancy within each categorical variables and ordered it by the difference. This Series was helpful later in the fitting process of the linear regression.

There is only one numerical feature variable in the dataset, the number of rooms in the possession of the household. It turned out the correlation is rather weak with a Pearson's r of only about 0.186. The p-value of the correlation is 0 however, so the correlation is statistically significant.

I now had analyzed whether or not the feature variables are statistically significantly related to the target variable. But for the purpose of creating a linear regression model I also needed to know how each feature variable is related to each other feature variables. There can't be correlations between categorical variables, but instead, there can be association. As a statistic for association, I've chosen Cramer's v. For each possible pair of categorical variables, I calculated the Cramer's v. A Cramer's v of 0.3 or more is generally considered as signifying a strong association. In order to get an overview of the association situation, I plotted a heatmap of the Cramer's v of each pair of categorical variables. Furthermore, I created a network plot (which can be seen below). Each node is a categorical variable and each connection means that there is a strong association between the connected variables. Categorical variables that are not strongly associated with any other categorical variable are not present in the plot). For using it later for the feature selection for the linear regression model I've also created a dictionary where each key is a categorical variable and the values are lists of categorical variables strongly associated with the respective key variable.



Now I had to find out if the only discrete variable, the number of rooms in the possession of the household, is statistically significantly correlated with any of the categorical variables. As there are strong outliers in this variable, I’ve chosen the spearman’s ranked order correlation. It turned out, this variable isn’t strongly correlated with any of the other feature variables.

Fitting the linear regression

For the linear regression, I one-hot encoded all categorical variables. The only numerical values I rescaled to have values between 0 and 1. As the scoring metric for evaluating which model is the best, I've decided to use Mean Absolute Error. The reason for that is, that I wanted to have an easily explainable metric and even more importantly a metric that is not strongly influenced by residual outliers, which are inevitable when predicting life expectancy. As a loss function, I decided to use the mean squared error.

For each iteration of the fitting process, my function made a scatter plot of the predictions versus the residuals, a histogram of the residuals, a quantile-quantile plot of the residuals, and a scatter plot of the predicted values versus the actual values. Furthermore, statistics such as the correlation between the predicted values and actual values, the p-value of the D'Agnostinos and Pearson Normaltest for testing the normality of the residuals, the mean of the residuals and the Mean Absolute Error were calculated. In order to prevent overfitting, I've always used 3-fold cross-validation.

Before making further adjustments I tried to find the right mix of feature variables by adding step by step more variables and assessing whether or not they improved the model. The order of adding new variables was determined by the order of the biggest difference in life expectancy of the categories within the variables. Variables were left out in case there was already another variable in the model to which it was highly associated. Of course, I also added the only one numerical variable to the model. It turned out that the model improved with each added variable, but only very slightly. So the best model so far was the one with most feature variables. I drew the following conclusions about this model:

- It did not perform very well, the mean absolute error was about 11.32
- the predictions-residuals scatter plot showed that one of the assumptions of linear regression was not met: There was Heteroscedasticity.
- Another not met assumption: The residuals were not normally distributed
- The correlation between actual values and predicted values was very weak, Pearson's r was only about 0.24.
- There were no predictions above 34.6 years and only a few above about 70

Next, I tried how the model performed leaving out rows that are outliers in terms of life expectancy. At the first view, the performance of this model was slightly better, due to the slightly lower Mean Absolute Error. But as the correlation between actual values and predicted values was weaker, I concluded that the lower Mean Absolute Error was most likely not due to better performance, but only because more extreme actual values under random circumstances lead to higher residuals. For that reason, I readded the outliers to the model.

Due to how the actual values versus predicted values scatter plot looked, I assumed that fitting without an intercept could perform better. But trying it out showed that this was not true. I've also tried converting the target values by taking the log, which didn't help as well.

At this point, I considered using another regression model than simple linear regression. The fact that the cross-validation score improved when adding more variables to the model and that all predictions were above about 34.6 made me assume that bias was the main contributor to the bad performance of the model rather than variance. Because the predictors obviously were not too strong but rather not powerful enough, I decided that regularized linear regression such as ridge regression or lasso is not an option.

The non-linearity of the residuals of the linear regression model signified that a non-parametric model such as for example Kernel Regression might be a good option. But unfortunately, my hardware power if far too weak for trying out such a model with even a sample of this data in a realistic time frame.

I decided to finally fit the whole training set and predict with the test set. The Mean Squared Error was 11.45 years. This is only a slight bit worse than the one I've got by cross-validating using only the training set. This is another proof for my hypothesis that the bad performance of the model is due to bias and not due to overfitting. Probably this data just doesn't allow for a better score. Next, I analyzed the coefficients of the model. The result of this helped me to formulate my recommendations. But before I'll explain those I will describe the additional last step I've done.

‘Experimental’ linear regression with personal data

The main part of the project was over now. But as the model didn’t perform very well, I needed to make a different recommendation for my client. As I’ve explained before, in the cleaning process I’ve had have to delete all variables about the deceased persons themselves, because the survey was set up in a faulty way. If it would have been performed in the right way, I could have used those variables for fitting the data. Perhaps it would make sense to repeat the survey in a better way. Of course, this would take on more resources and the results could only be used in future decades because the client would have to wait until the surveyed die to know at which age they died. On the other hand, it could be useful to have a better model. So I supported my client making this decision by making an experiment.

The experiment was readding the personal variables and again performing linear regression with the additional data. It's important to note that this ‘experimental' model can't be used for real predictions because it contains the faulty variables. It also can't be used to precisely assess how the model would perform when redoing the survey in a correct way and using the personal variables, because those variables would not be faulty. But it can give us an intuition how well the model MIGHT perform with more variables.

Before making the fitting I had to do parts of the Data Cleaning, Exploratory Data Analysis, and feature selection one more time with the new data. I won't describe this process further but jump to the outcomes. The ‘experimental' linear regression was still not performing very well, but clearly better than the ‘real' model. It performed better in the following ways:

- The Mean Absolute Error improved by about 2 years.
- There is far less Heteroscedasticity
- The Pearson’s r of the ‘real’ model is about 0.23, the ‘experimental’ Pearson’s r is about 0.6.

Recommendations

The recommendations are the following:

- **As the model does not perform very well I would use it only with care. Important decisions should not be based solely on my model. I wouldn’t base the allocation of resources on the life expectancy as predicted by this model.**

- **The model can, however, be used to get the first impression when searching for households and individuals who need special attention, or can be helpful for further research. Furthermore, staff on the field can be thought to use the coefficients of the model for finding households that have an especially high or low life expectancy. Good indicators for a life expectancy are for example religion, the toilet the household uses or the number of rooms in the possession of the household.**

- **I would consider redoing the survey in order to be able to create a better model with more variables. Such a survey should be done in a way that creates a dataset in which the observations are individuals either at a fixed specific age or at a random age, instead of at the time the person dies. This would allow for using variables about the persons themselves. It can't be said with certainty that this would improve the model, but judging by the ‘experimental' model I've done it could.**