Virtual Data Analytics Internship May 2023

Data Processing, Exploratory Data Analysis and Machine Learning

Contents

Important information

1. The deadline for this project is three weeks from acceptance.
2. Your report can be written in R Markdown (Rmd, if R) or Quarto (if Python).
3. The page upper-limit for your report is 12 pages of text, excluding plots. There is no page limit for your code.
4. Your report and R/code file must be submitted to your mentor's mail by 9pm of due date.

Task description:

This project aims to bring all the skills you have learnt and give you an opportunity to apply them to a dataset. Such a dataset can be chosen by yourself, which can be relevant to financial, insurance, medical or health areas etc. We provide the Ames Housing dataset (an American housing dataset) if you would prefer it. A brief description of the variables in the dataset is at the end of this file.

In this project, you will need to write a full report on your analysis of your chosen dataset, from the beginning of the data science methodology, where you will need to establish your problems of interest/exploration to the end of the *Further Pre-processing* stage. You will then train a simple linear model on the training dataset and predict values for the *test* dataset. Finally, you will evaluate your model using the metric RMSE (if regression problem) against the *test* dataset and plot the residuals, and draw your final conclusions.

Your report is to obtain insights and aim to make impacts, and should have a structure which follows the data science methodology. The main purpose of this project is to conduct your analysis using EDA and visualisation and use DevOps tools like git/github for version control and reproducibility.

For example, on the housing dataset (if chosen by you), when writing the report, put yourself into the shoes of a real estate analyst wanting to obtain insights from this dataset to predict house prices. The dataset already has a lot of reports written on it – find them here. Be inspired by them for EDA, but do not focus too much on their modelling.

Use the datasets labelled *train* and *test* . As you are a real estate analyst, your target variable is *SalePrice*. Note for most of the report, you will only use the *train* dataset. This includes preprocessing, EDA, and everything else up to and including the creation of a linear model.

The linear model will then be trained on the *train* dataset. You will then predict a set of SalePrice values based on the variable information in the *test* dataset. You can then compare your predicted values to the 'real' values in the test dataset. Therefore the *test* dataset is only needed for the "Evaluation" section of the report.

Report structure

Your report needs to include the following sections. In each section you will need to give a very brief explanation as to what the section is about, what the purpose of the section is and/or describe the key pieces of information in your general approach. For example, in the "Data pre-processing" section, you would explain what exactly data pre-processing is, why you need to clean the data, and describe the key ideas in your approach e.g. fill in missing values with median based of external controls.

Title and abstract:

On the first page, you should have:

- A suitable title for your report
- Your name
- An abstract/executive summary outlining your problems, analysis and findings.

Problem identification:

You should conduct some background research into the *Ames Housing dataset* and:

- Give some information on the dataset.
- Gather and list points of domain expertise to help you make better decisions and shape your report (e.g. you should identify creating a variable similar to Week 7/9's *SeasonSold* would require you to know which seasons correspond to which months as the dataset is American)
- Seek to understand the variables here.

Problem identification and understanding is crucial in any data science project. You should:

- Think about (after gaining domain expertise) a few questions of interest, which you will then translate into data science problems to solve within your report (if you get stuck look at a few examples from the Melbourne dataset slides with problems of interest).
- Provide a list of these data science problems. You will need to address and interpret your corresponding findings later on in the body of your report.

Note that **examples of problems** for you to find and solve can be:

- Identify which suburb/location had the biggest growth in *SalePrice* by plotting and examining the sale prices cross different suburbs;
- Analyse a possible pattern of *SalePrice* vs *YrSold*/*MoSold*, *LotArea* and/or some other variables which can reasonably be included;
- Use predictions from your final model to compare suburbs which have shown varying growth. Or, to identify which suburbs have been growing the most over the last few years.

Data pre-processing:

In this section you should:

- Pre-process your code, treat missing values etc.
- Note at least one key observation, e.g. identified possible missing values or outliers for a particular area/suburb or year e.g. 2016 is significantly higher. Or perhaps one column is missing more than 50% of its values.

EDA:

In this section you should:

- Include tasks such as determining which variables are significant, which observations may be outliers etc., and other EDA goals.
- Find as much insight as possible to support your modelling decisions later on.
- Use data visualisation techniques taught in the unit to answer your chosen problems of interest.

Further pre-processing:

In this section you should:

- Select the final variables for your model based off your EDA (basically remove the non-significant variables).
- Create any new variables which you think may help based on your EDA in this section.
- Justify your decisions and provide EDA evidence as to how a variable is insignificant (e.g. no observable relationship to target variable in scatter plot).

Modelling:

In this section you should:

- Fit and evaluate a linear model to describe the relationship between your target variable and a number of selected significant predictors.
- Use your model to predict the prices of properties described by your *test* dataset.

Alternatively, you may use another, more advanced model of your choice. If you do use a linear model, remember its likings such as a normalised distribution in the target variable.

Evaluation:

You should:

- Evaluate your model against the metric RMSE given the actual values in the ***test* dataset**
- Plot the residuals. Pick a suitable cut off value for the red dots.

The data science methodology is an iterative process. Try to minimise your RMSE, so always go back and think about what improvements can be made, then fit another model, and find your second RMSE, and so on, noting what works and what does not. Compare at least two different models you considered, noting their differences.

Recommendations and final conclusions:

You should:

- Summarise your findings and provide your found solutions to your problems of interest. Note anything you found particularly interesting and useful to your project.

- State the best **RMSE** you obtained and why/how (i.e. what variables you used, any applied transformations etc.).
- State any improvements you could make and why/how you could achieve such improvements in future works.

References:

You should:

- Include a reference list and cite your references via in-text referencing or footnotes.

Variables in the Ames Housing dataset:

Below, please find a brief description of the variables within the dataset. For more detail, look inside the *data_description.txt* file.

- *SalePrice* - the property's sale price in dollars. This is the target variable that you're trying to predict.
- *MSSubClass*: The building class
- *MSZoning*: The general zoning classification
- *LotFrontage*: Linear feet of street connected to property
- *LotArea*: Lot size in square feet
- *Street*: Type of road access
- *Alley*: Type of alley access
- *LotShape*: General shape of property
- *LandContour*: Flatness of the property
- *Utilities*: Type of utilities available
- *LotConfig*: Lot configuration
- *LandSlope*: Slope of property
- *Neighborhood*: Physical locations within Ames city limits
- *Condition1*: Proximity to main road or railroad
- *Condition2*: Proximity to main road or railroad (if a second is present)
- *BldgType*: Type of dwelling
- *HouseStyle*: Style of dwelling
- *OverallQual*: Overall material and finish quality
- *OverallCond*: Overall condition rating
- *YearBuilt*: Original construction date

- *YearRemodAdd*: Remodel date
- *RoofStyle*: Type of roof
- *RoofMatl*: Roof material
- *Exterior1st*: Exterior covering on house
- *Exterior2nd*: Exterior covering on house (if more than one material)
- *MasVnrType*: Masonry veneer type
- *MasVnrArea*: Masonry veneer area in square feet
- *ExterQual*: Exterior material quality
- *ExterCond*: Present condition of the material on the exterior
- *Foundation*: Type of foundation
- *BsmtQual*: Height of the basement
- *BsmtCond*: General condition of the basement
- *BsmtExposure*: Walkout or garden level basement walls
- *BsmtFinType1*: Quality of basement finished area
- *BsmtFinSF1*: Type 1 finished square feet
- *BsmtFinType2*: Quality of second finished area (if present)
- *BsmtFinSF2*: Type 2 finished square feet
- *BsmtUnfSF*: Unfinished square feet of basement area
- *TotalBsmtSF*: Total square feet of basement area
- *Heating*: Type of heating
- *HeatingQC*: Heating quality and condition
- *CentralAir*: Central air conditioning
- *Electrical*: Electrical system
- *1stFlrSF*: First Floor square feet
- *2ndFlrSF*: Second floor square feet
- *LowQualFinSF*: Low quality finished square feet (all floors)
- *GrLivArea*: Above grade (ground) living area square feet
- *BsmtFullBath*: Basement full bathrooms
- *BsmtHalfBath*: Basement half bathrooms
- *FullBath*: Full bathrooms above grade
- *HalfBath*: Half baths above grade
- *Bedroom*: Number of bedrooms above basement level
- *Kitchen*: Number of kitchens
- *KitchenQual*: Kitchen quality
- *TotRmsAbvGrd*: Total rooms above grade (does not include bathrooms)

- *Functional*: Home functionality rating
- *Fireplaces*: Number of fireplaces
- *FireplaceQu*: Fireplace quality
- *GarageType*: Garage location
- *GarageYrBlt*: Year garage was built
- *GarageFinish*: Interior finish of the garage
- *GarageCars*: Size of garage in car capacity
- *GarageArea*: Size of garage in square feet
- *GarageQual*: Garage quality
- *GarageCond*: Garage condition
- *PavedDrive*: Paved driveway
- *WoodDeckSF*: Wood deck area in square feet
- *OpenPorchSF*: Open porch area in square feet
- *EnclosedPorch*: Enclosed porch area in square feet
- *3SsnPorch*: Three season porch area in square feet
- *ScreenPorch*: Screen porch area in square feet
- *PoolArea*: Pool area in square feet
- *PoolQC*: Pool quality
- *Fence*: Fence quality
- *MiscFeature*: Miscellaneous feature not covered in other categories
- *MiscVal*: $Value of miscellaneous feature
- *MoSold*: Month Sold
- *YrSold*: Year Sold
- *SaleType*: Type of sale
- *SaleCondition*: Condition of sale

References:

- Datasets can be downloaded from here - https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data
- Inspiration (other reports on this dataset) - https://www.kaggle.com/c/house-prices-advanced-regression-techniques/notebooks?sortBy=hotness&group=everyone&pageSize=20&competitionId=5407&langu age=R