International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 6 | ISSN : 2456-3307

Capstone Projects: Data Science in Real Life

Pranav Murali
SRM University, Chennai, Tamilnadu, India

ABSTRACT

This is a paper showcasing two capstone projects that will help us understand about data science and it's role in a real life corporate. The first capstone project was initiated by John Hopkins University as part of a specialisation course. It was a project in collaboration with an industry partner named "Zillow" , an online real-estate, rental website in the United States. The second capstone project was initiated by University of California Davis, also as part of a specialisation course. It was a project in collaboration with an industry partner named "Tableau", an online software company that is involved with business intelligence and analytics.

Keywords : *Baltimore, Data Scientist , Sales , Hypothesis , Model Validation , Overfitting , Importing , Visualisation*

I. INTRODUCTION

Part 1 : Executive Data Science

It involved a group of data scientists, managers and data engineers. I was given a brief introduction about what each data engineer, data manager and data scientist of the company would be doing in my project and their specific roles which would help me . The project details about Zillow, an online real estate, rental website were given to me and I was presented with a data set about the city of Maryland, Baltimore with various details about rental houses and houses on sale. Details such as square footage, year of construction, no. of bedrooms, no. of baths and so on were listed. The issue with the data was that there were some missing values for many houses and it needed to be addressed immediately so that it could be properly uploaded in their website. I was given 3 choices: to impute the missing values or build a predictive model or use only properties with adequate data. Since the problem was missing data and not about manipulation of data from a database for presentation to the managers, this project mainly involved the work of a data scientist. I had to make a decision of the three choices and imputing the missing values was the logical one.

Data and Statistics for the project:

Once the decision was made regarding the choices, I was given the summary statistics consisting of property map, sales , square footage , area details , year of construction and so on.

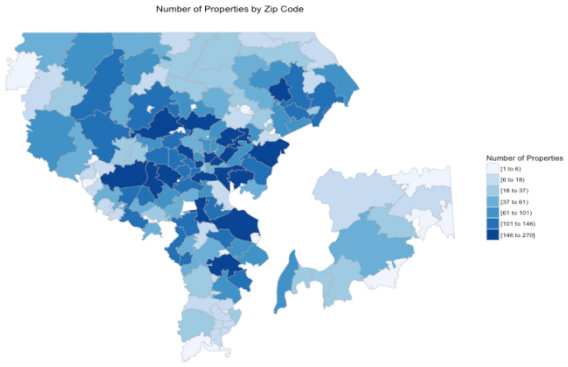

Figure 1. No. of properties by zipcode

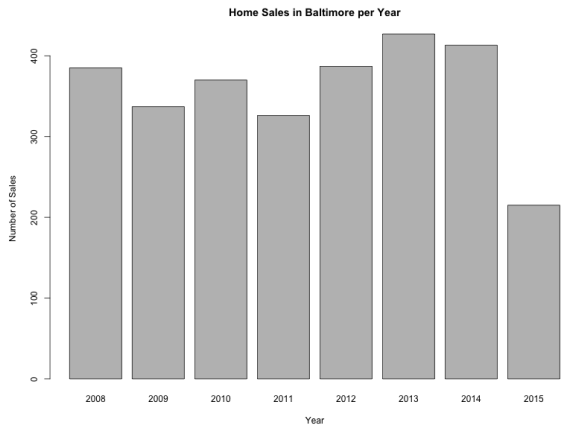

Figure 2. Sales in Baltimore per year

Figure 3. Year of Construction

Incorporating the data:

Once all the data, statistics for the Baltimore state was given by the company, I had to find a feasible way to incorporate them in a model and bring up the model to rectify the missing data and update the files with proper details about all the houses in the Baltimore state, which was the goal of the project.

Incorporating the data could be done in 2 ways:

- Overfitting/Cross Validation
- Model Validation

Overfitting/Cross Validation:

Let's say you attend a symphony and want to get the clearest, most faithful sound possible. So you buy a super-sensitive microphone and hearing aid to pick up all the sounds in the auditorium.

Then you start "overfitting," hearing the noise on top of the symphony. You hear your neighbors shuffling in their seats, the musicians turning their pages, and even the swishing of the conductor's coat jacket.

When you're at a concert, there's both the symphony and the random noise. Fitting a perfect model is only listening to the symphony. Over-fitting is when you hear more noise then you need to, or worse, letting the noise drown out the symphony.

Talking about this in a more technical sense, **in any dataset we have the signal and the noise.**

Let's say we're trying to predict sales in December for successful retailer. The signal is how seasonality, upwards trends, and consumer sentiment affects sales numbers. The noise is just random variation that occurs that doesn't have anything to do with our predictors.

When we do prediction, we want to identify the signal. We want to identify what the seasonality effect is. We want to identify how consumer sentiment effects our sales. We want to understand our upwards trend in sales. There's no use in trying to identify the noise since the noise will change every month.

Our predictions will be the most accurate if we can model as much signal as possible and as little noise as possible. This is why R^2 is a bad metric to use to identify predictive power - R^2 measures how much of the signal AND the noise is explained by the model.

Unfortunately, it's hard to always identify what's signal and what's noise. This is why practical applications tends to favor simpler models, since the more complicated a model is the easier it is to fit to noise. Quantitative hedge funds rely on very simple techniques like regressions and shun complicated techniques (like many ML algorithms), to try to achieve understandable models that have the highest prediction capability.

Model Validation:

Many techniques exist to combat model over fitting. The simplest method is to split your dataset into training, testing and validation sets. The training data is used to construct the model. The model constructed with the training data is then evaluated with the testing data. The performance of the model against the testing set is used to further reduce model error. This indirectly includes the testing data within model construction, helping to reduce model over fit. Finally, the model is evaluated on the validation data to assess how well the model generalizes.

A few methods where the data is split into training and testing sets include : k-fold cross-validation, Leave-One-Out cross-validation, bootstrap methods, and re-sampling methods. Leave-One-Out cross-validation can be used to get a sense of ideal model performance over the training set. A sample is selected from the data to act as the testing sample and the model is trained on the rest of the data. The error on the test sample is calculated and saved, and the sample is returned to the dataset. A different sample is then selected and the process is repeated. This continues until all samples in the testing set have been used. The average error over the testing examples gives a measure of the model's error.

II. CONCLUSION

Finally, with the assignment of proper variables in the model, a hypothesis was made and the regression model for the missing data was developed. Thus, Baltimore state in the website was updated with proper details.

Part 2 : Data Visualisation Using Tableau

Develop a Project Proposal:

A project proposal that will capture the "who, what, why and how" of the project . The proposal included: a specific business case or personal objective, any intended outcomes, a description of the needs of the intended audience, a description of the dataset to be used, and any foreseeable challenges.

Details:

This work was regarding the mobile phone usage in a region. Data regarding the percentage, no. of users were given. A graphical analysis was done to determine the leading mobile phone company.

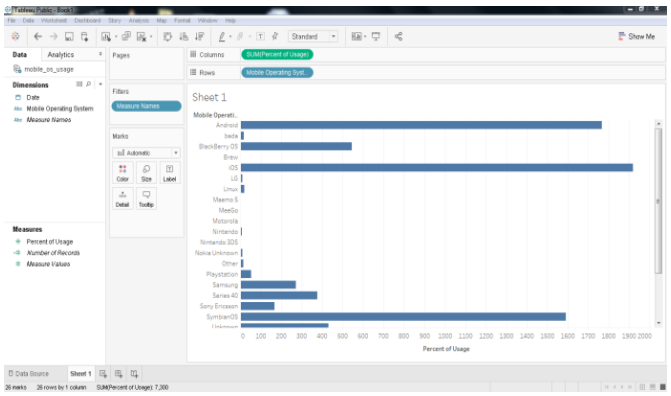

Figure 4. Bar graph analysis of the mobile phone usage in a region using TABLEAU

Importing and Prepping the Data:

In this step, the dataset that supports the project proposal was imported it into Tableau, and prepared for analysis.

Data Analysis:

Once the data was imported into tableau work space, explanatory and exploratory analysis had to be done. KPIs were created (Key Parametric Indexes) and visualisation was done.

Actions, filters, hierarchies were used in the process of visualisation. Finally, dashboards were created to finish the structure.

Storytelling and Presentation:

Last part of the project was presenting the data with visualisations and communicating the details in the right way.
Presentation is an important part because that only decides the feedback part for the work that you have done.

III. Conclusion

It started with data preparation, then came importing the data and then analysis of data both explanatory and exploratory. Finally, graphical visualisation and presentation of the project was done.

IV. REFERENCES

[1]. https://capstone.datacamp.com/courses/executive-data-science-capstone-project/capstone-project-project-title?ex=11

[2]. https://github.com/booz-allen-hamilton/The-Field-Guide-to-Data-Science/blob/master/LifeInTheTrenches-LessonsLearned/ModelValidation/ModelValidation.md

[3]. https://www.quora.com/What-is-an-intuitive-explanation-of-over-fitting-particularly-with-a-small-sample-set-What-are-you-essentially-doing-by-over-fitting-How-does-the-over-promise-of-a-high-R%C2%B2-low-standard-error-occur/answer/William-Chen-6?share=6120088f&srid=O6WD

[4]. https://www.perceptualedge.com/blog/?p=1897