Erdi Olmezogullari – Kaan Dandin
Data Mining – Progress Report

First Glance to Data

We retrieved Amazon Spot Price data which was collected by a third party person. Data was basically separated to epoch by day-wise. Each epoch has 5 fields <Timestamp, ProductDescription, InstanceType, SpotPrice, AvailabilityZone> in JSON format. At the beginning of the data mining process, we wrote a simple shell script in Linux so that we transform the data to CSV format. According to data fields, Timestamp (TS) is a time when the data was collected. Product Description (PD) is referring to kind of operating system on an instance of Virtual Machine (VM) where Operating System will be installed according to customer's demands. It is consisting of 6 unique operating systems. Instance type (IT) is referring to type of VM. Since IT can be picked with respect to business' goal of customer, it was taken down into wide brands with 33 unique VM types by AWS. Spot Price (SP) is showing the current market price for each IT and Availability Zone (AZ). AZ is consisting of 22 unique zones in different countries across the world.

Exploring Data

After data transformation, to take a quick look at the data by using simple decision tree by using Rapidminer, we tried to build our first primitive decision model over the formatted and cleaned data. According to this quick overview, we decided that AZ will represent a labeled attribute of our decision tree. Since it has 22 different zones, it would not represent as a binominal attribute. In this phase, we consider that it would not useful information to tag the attribute on as a polynomial because of fact that we did not get the ROC curve of the model for polynomial labelled data instead binominal. Our final decision is to define the attribute is as a binominal attribute. To perform this methodology, we defined our strategy to process data by using decision tree. In our strategy, after every AZ was once tagged as a True (or its name), the rest of them were tagged as a False (or "Other"). Therefore, since we have 22 different zones, we generated 22 different files based on each AZ after we apply this strategy on the one file which belongs to different AZ for a day.

Future Works

In future work, we will build 22 different models for each file which we generated by using AZ (True, False or Name, Other). In the next stage, we will compare those 22 different models by using ROC curve. According to this comparison, we will expect to find a model which will be having min error and high accurate decision. At the end of this process, a final ROC curve will represent just one day. To compare the rest daily deviation, we will process the rest of day-wise files. After those processing, we expect to find the best model within 30 days. But, perhaps, it will be neither a general nor distinct model if the models are having a deviation among each epoch. In this work, we will observe and compare treading of whole models during 30 days, too. According to these results, a few models may cover either whole days or not.