Exploratory data analysis

To choose features one could use:

1. Filter Methods: (to choose features)
    a. Chi squared test
    b. information gain
    c. correlation coefficient scores
2. Embedded Methods (to choose features during “construction” of model)
    a. LASSO
    b. Elastic Net
    c. Ridge Regression
3. Wrapper Methods (searching through the space of subsets of features)
    a. Best-first search
    b. Random forrest algorithm
    c. Random hill-climbing algorithm (love this one)
    d. Recursive feature elimination algorithm.

We want to:

- Find information about every feature and the data set as a whole
    - Data type
    - Variance
    - Outliers or extremes? Search through and see what you find!

Data preprocessing

Cleaning:

- What needs to be cleaned:
    - -999 values would fuck up the mean, variance etc.. (these are values that could not be measured).
    - There is 10 columns that contain a lof of -999, and one column that contains a lot of 0’s. These columns need special treatment.

Feature Processing

Ranking features and picking subset

To rank features one could consider the following attributes:

- Check the covariance between the prediction and the feature.
    - Each attribute is considered independently

Selecting features while constructing the model (probably not usefull for us)

More complex predictive modeling algorithms perform feature importance and selection internally while constructing their model.

- MARS
- Random Forest

- Gradient Boosted Machines.

Feature Extraction (creating features) :

- The automatic construction of new features from raw data
- **PCA**

Other methods for Feature selection

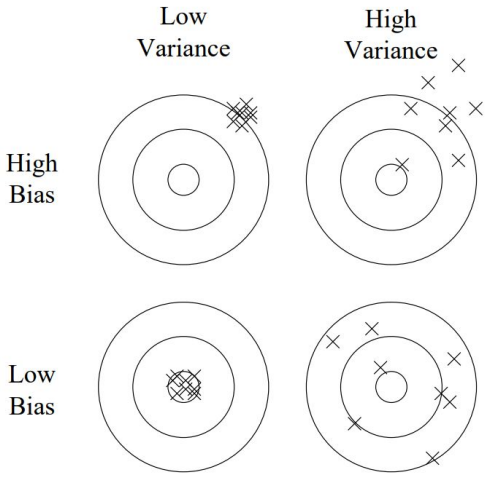
- Choose (fully/partly) random subsets of the features, and test how well they do.

We want to:

- Select features to use.
- Make new features
    - EXCEL has a LOT of tools we can use here, but it's manual work, requires brain power!
    - Check out this link: https://www.youtube.com/watch?v=drUToKxEAUA
- Proposals to methods we can use
    - Rank the features by covariance with the prediction
        - Choose top X features, could test solution for 1 - 20 features.
    - Use PCA to extract (new) features (might be a bit technical)
        - Choose top X features, could test solution for 1-20 features.
    - Take a random subset of features, run it on small subset of training data. Run A LOT of random subsets, choose the best one.
- When you find a good feature:
    - Find similar features, and combine them into a simpler representation
    - If it's a CUT OF feature, say it's 1 for values over 30 and 0 for values less then 30, try to iterate over different values for the cut of, to see which is better.

Evaluation

- If it's not too computationally heavy
    - Cross validation if it's not to much calculation work
- If it's heavy compuntationally
    - Deviding the test-data into two parts, train and test. Ish 70 / 30 %



Masterplan

Clean data

1. There is 10 columns that contain a lof of -999, and one column that contains a lot of 0's. Remember that these columns might need special treatment?
2. Count the number of -999 and 0's in each feature. Interesting?
3. Remove -999, set to 0. (? not sure if that's a good solution ?)

Preprocessing

1. Normalize features with mean normalization
   https://en.wikipedia.org/wiki/Feature_scaling

Feature processing

1. Process features that can be split, enhanced. Make new ones!

..... (CONTINUE WORK HERE) .....

2. Find similar features, and combine them into a simpler representation
3. If it's a CUT OF feature, say it's 1 for values over 30 and 0 for values less then 30, try to iterate over different values for the cut of, to see which is better.