DEMO BY DEMO

Chapter A
Feature Selection

Loaii abdalslam

هنتكلم دلوقتي عن ال Feature Selction اول لما تقابلك الداتا فأنت بتقابل مشكلة وهي تختار انهى Feature وليه تختار ال Feature دي وامتا تختارها , وهل طرق الاختيار تختلف بناء علي الداتا ونوعها وتفاصيلها ولا لأ الشابتر ده هنتكلم عن كل تفاصيل ال Feature Selection وازاي نقدر نتعامل معاه .

في المشاكل الصغير مش بتكون عامل حسابك على Processing Power الي بتستخدمها لل Modeling كل الي بيكون في بالك انك توصل لأعلي ACCURACY , وطريقة الشغل دي مش هتكون مناسبة لو هنقرب جنب ال Big Data لأن طريقة الشغل وقتها هتختلف لاننا هنكون محتاجين نقلل ال Power دي لأن ال Power دي بتحتاج فلوس او بتحتاج موارد أعلى ومش في كل الحالات الشركة او المؤسسة او إمكانياتك بتوفر ده علشان كده بنلجأ لل Feature Selection وهو بكل بساطة اني اختار من ضمن ال Feature الموجودة عندي أحسن Feature هتفيد عملية ال Modeling وفي أكثر الحالات يلجأ لبعض ال Functions الي تخلينا نقلل ال Feature دي .

ال Features الى بنختارها بتكون أهم Features موجودة عندنا وهي فعلا التغير فيها يؤثر علي الناتج النهائي لل Model ومعدل الأداء .

النوع الأول ال Filter Based :

النوع ده قائم على المعادلات الإحصائية التي تبين مدى الترابط بين ال input values وبعضها , الفكرة كلها ان كل ال Independent variable المفروض تكون مستقلة لأن ال Target Value بتاعتنا الي المفروض اننا نتنبأ بها بتكون اعتمادية - Depended علي ال Feature المستقلة , طيب ولنفترض ان معامل الارتباط لبعض من ال Feature مرتفع وإحنا نستغني عنه تماما وكثير من الاحيان بتكون قيمة الترابط Correlation بتكون عالية جدا وممكن ناخد مثال علي الجزئية دي

مثال 1 :

تخيل معايا ان سعر البيت بيتحدد على مجموعة من ال Features (مساحة المنزل - مساحة الجراج - عدد العربيات التي تتسع للجراج)
بمجرد النظر لأسماء الى Columns هتكون عارف ان عدد العربيات التي تتسع للجراج ترتبط ارتباط مباشر بمساحة الجراج , فكلما ارتفع معامل الارتباط بين متغيرين فداه معناه أننا نستبعد واحد منهم تماما من ال Modeling تماما وتعتمد على Feature واحدة وهي هتكون مساحة - مساحة المنزل علشان نتنبأ السعر .

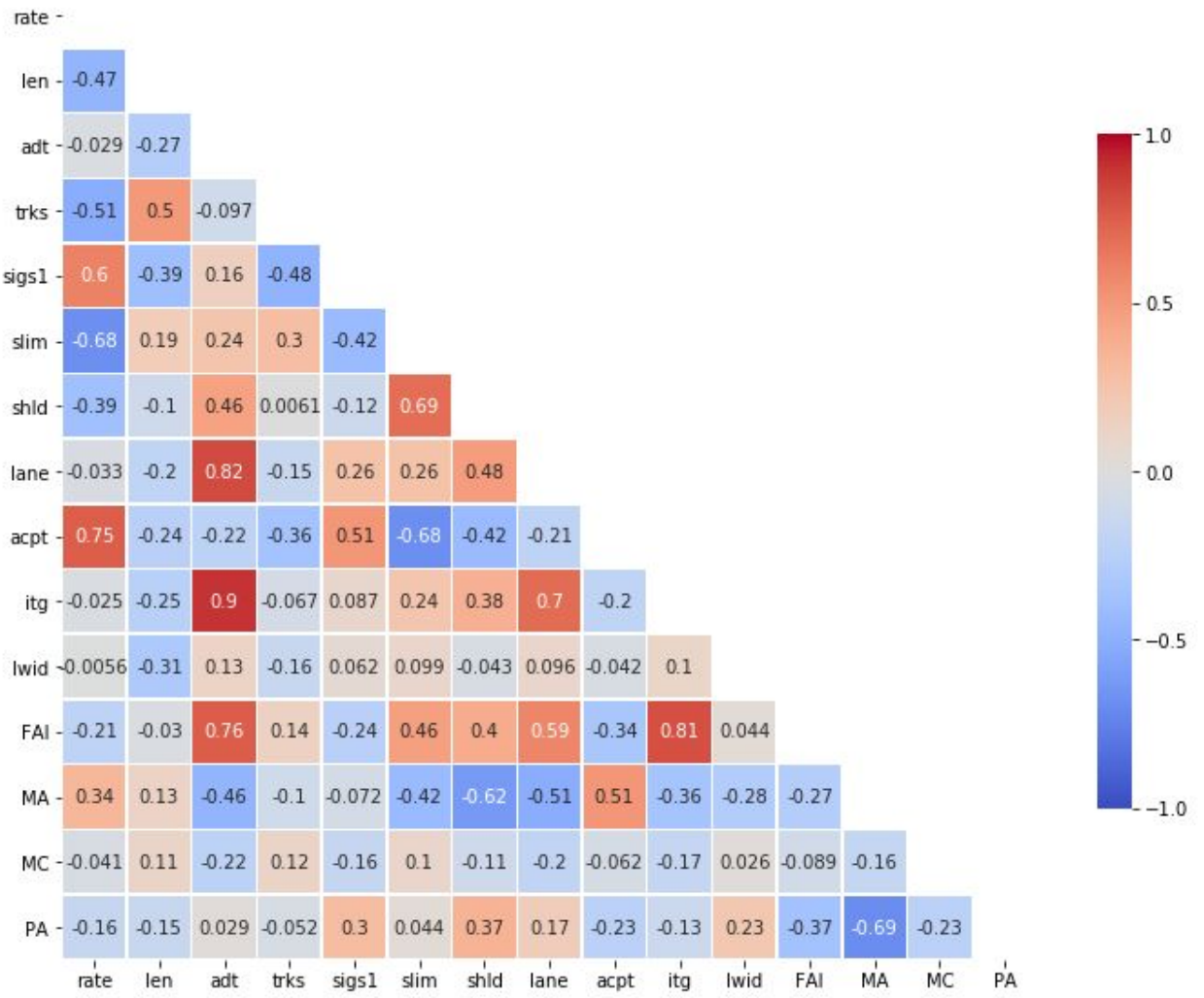
$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2\,(n\sum y^2 - (\sum y)^2}}$$

$$r = \frac{\sum Xy - n\,\bar{x}\,\bar{y})}{\sqrt{(\sum x^2 - n\,\bar{x}^2)(\sum y^2 - n\,\bar{y}^2)}}$$

معامل الارتباط لبيرسون

مثال 2 :

ولنفترض انك معاك Heat Map مثل الموجودة معانا هنا , ازاي تقدر تقراها علشان تطلع ال Correlation الموجود بين ال Feature , لو ذاكرت مادة معاملات الارتباط فأنت من المفترض تكون عارف ان كل ما بنقرب لل 1 فداه معناه ان معامل الارتباط اعلي ولو قربنا لل 0 فداه معناه ان مفيش ارتباط ومش بالضرورة يكون مفيش ارتباط لأن احيانا الإرتباط بيكون لا تمثله علاقة خطية , بالإضافة ان فيه إرتباط عكسي وده بيكون بسبب يعني , يعني مع زيادة X1 في المقابل X2 هتقل وهكذا ..



دلوقتي انا هعتبر ان Threshold range بتاعي إن ال Feature بتاعتي موجودة في $F \in [-0.6, 0.6]$ وما دون ذلك يعتبر اعتمادي على بعضه , فلو أخذنا على سبيل المثال عينة عشوائية من ال High Correlated Feature هنلاقي إن :

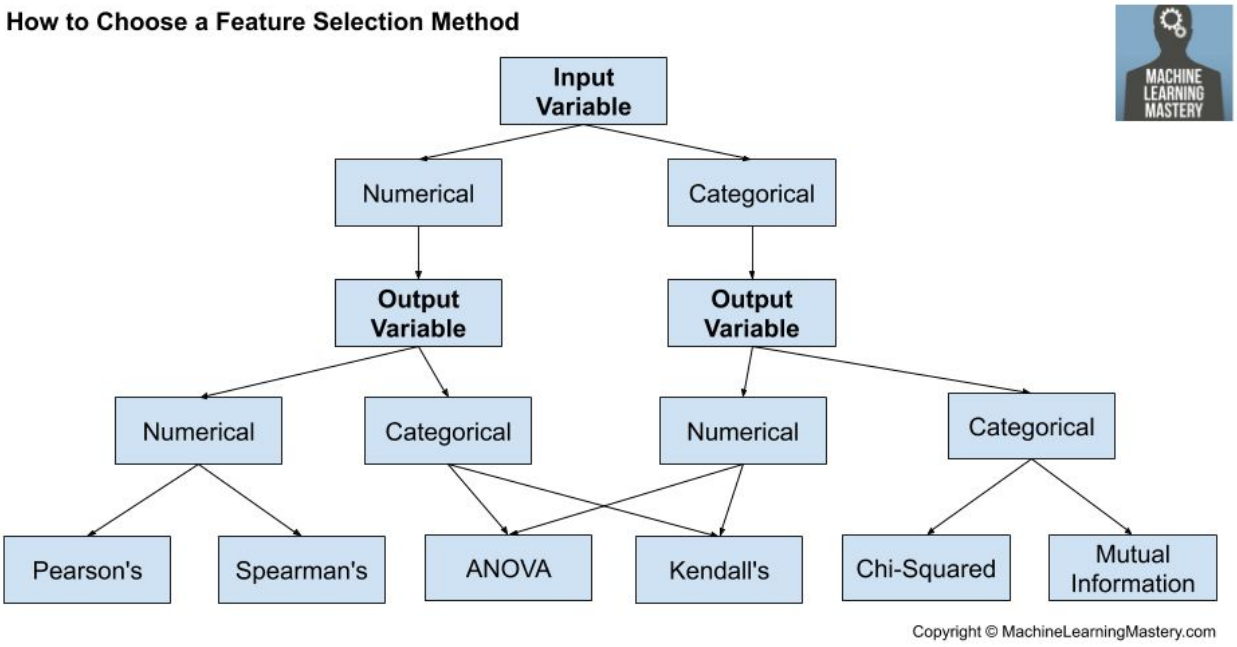
```
adt ~ itg = 0.9
adt ~ FAI = 0.76
adt ~ dat = 0.82
```

نلاحظ ان ال adt مقارنة بباقي ال Feature الي هي مرتبطة معاهم أنها مرتبطة مع أكتر من حاجة فالأولية لينا اننا نشيلها خالص من ال Model بتاعنا متعتمدش عليها تماما ونعيد تكرار العملية لحد ما نشيل كل ال High-Correlated Features , وبكده نكون حلينا مشكلة الاعتمادية في Feature 4 ونشوف Feature غيرها علشان نبحث فيها وهكذا , وطبعا معامل الإرتباط يعتبر واحد من ضمن Functions كثيرة نشرحها في المستقبل القريب .

النوع الثاني ال Wrapper Based :

الطريقة دي عكس الطريقة الأولي تماما الفكرة كلها اننا هناخد حبة Features بشكل عشوائي ونحاول اننا نعمل بيهم Model ونشوف احسن Model ونبدأ نبدل ال Features علشان نشوف كام Feature بتأثر تأثير ايجابي علي ال Model وعلي مستوي الأداء بتاعه وال Performance Metric ايا كانت , لو هنستخدم Function مثلا شبه الي عندنا هنا اسمها RFE فهي وظيفتها انها تعمل Train علي كل ال Feature بتاعتك وتبدأ تقلل فيهم بحيث انها تاخد الأفضل فالأفضل لحد لما توصل ل أفضل Features وتستخدمهم .

السؤال الي ممكن يخطر علي بالك دلوقتي وهو امتى بالظبط اختار نوع ال Feature Selection Technique على سبيل المثال ممكن تشوف الصورة الجاية علشان تحاول توضح لك وتبسط لك أفكارك



ملحوظة : كل اسماء ال Functions موجودة في Skict-learn & Scipy حاول تذاكرهم للضرورة وهنحاول نغطيهم في الاجزاء القادمة

الحالات دي هي الموجودة عندنا وفيه بعض النصائح المفروض انك تاخد بالك منها وانت شغال بحيث تتعامل مع شتى انواع الداتا بكل سلاسة ومن غير أي صعوبات

تواجهك , في بعض ال Functions الموجودة بتحتاج TYPE - نوع معين من البيانات ويكون صعب عليك انت تشتغل لو الداتا بتاعتك فضلت على نفس ال Type بتاع بعض ال Feature فيها علشان كده ممكن تستعين بالـ Variable Transformation

تحويل المتغيرات

علي سبيل المثال لو عندك Variables اسمهم (افريقيا - اسيا - اوربا) فأنت ممكن تحوله انهم بدل ما كانوا Categorical Feature انهم يكونو ordinal Feature وبعد التحويل هيكونوا بالشكل ده (1- 2- 3) .

ممكن نستخدم ال numerical variable وتحوله ل Discrete على سبيل المثال ال BINS