Kunskapskontroll del2

Personlig del

Keikiet Pham

EC Utbildning

Del 2

202404

Innehållsförteckning

1 Datainsamling 3
2 Teoretiska frågor 4
3 Självutvärdering 6

1 Datainsamling

1. Vem du har arbetat i grupp med?

Grupp 1: Adrian, Jakob, Keikiet, Melissa, Robert

2. Hur har ni i gruppen arbetat tillsammans?

Dag 1: Initialt diskuterade vi mycket om hur mycket data och vilka variabler som skulle samlas in och vilken data som skulle samlas in. Här diskuterades mycket fram och tillbaka. Några gruppmedlemmar inklusive mig själv, föreslog en smalare datainsamling där vi begränsade oss till ett fåtal av de mest populära bildmodellerna, i hopp om att bygga en mer robust modell medan andra ville ha en bredare datainsamling då detta ansågs kunna ha ett större användningsområde. I slutändan landade vi i en kompromiss där vi skulle samla in data för Volvobilar, på så vis undviker vi att introducera prisvarianser mellan olika bilmärken samtidigt som att vi har ett flexibelt dataset man kan bygga flera olika business case:s runt. Vi bestämde oss även för att samla in 50 datapunkter vardera i olika bränsletyper till en början och sedan utvärdera om mer data ska insamlas eller om något ska ändras.

Dag 2: Dagen efter skulle vi ha ett nytt möte där jag sammanställt och presenterat datan som samlats (enkel EDA i excel), då framkom det att stor del av datan bestod av företagsbilsäljare vilket var mindre önskvärt (vi vill ha en jämn fördelning variablerna i vår datainsamling). Vi bestämde oss för att varje medlem skulle göra en Proof of concept-modell i R studio och om allt såg OK ut, samla in ytterligare 100 punkter på olika bränsletyper med mer vikt på privata försäljningsannonser. Var man inte nöjd med datakvaliteten under POC-modelleringen skulle det uppmärksammas till gruppen.

Ett nytt möte bokas in till dagen efter.

Dag 3: Här presenterades ny sammanställd data och vi diskuterade PoC-modelleringarna. De flesta var nöjda med datakvaliteten och vi diskuterade idéer och funderingar till uppgiften. Vi bestämde oss för att nöja oss med datainsamlingen på 750 punkter.

3. Vad var bra i grupparbetet och vad kan utvecklas?

Överlag är jag mycket nöjd över grupparbetet, samtliga medlemmar var engagerade i diskussionerna och i grupparbetet. Kommunikationen var god men kan alltid förbättras, effektiv kommunicering är viktigt och önskvärt, särskilt i grupparbeten.

4. Vad är dina styrkor och utvecklingsmöjligheter när du arbetar i grupp?

Jag anser att några av mina styrkor är att jag är inkluderande och uppmuntrar samtliga gruppmedlemmar till att vara involverade i grupparbetet.

Jag anpassar mig efter dynamiken och kan vara drivande när det behövs samtidigt som att jag låter andra ta plats när det gynnar gruppen. Är målinriktad och strävar hela tiden för att grupparbetet ska fortsätta framåt.

Ibland kan jag vara för dominant och ha sämre tålamod särskilt när jag känner att vi inte kommer framåt.

5. Finns det något du hade gjort annorlunda? Vad i sådana fall?

Inget direkt, det var ett litet grupparbete och på så vis förlåtande arbete. Kanske var jag för seriös och rättfram och skulle pratat mer om vädret? :). Tror att generellt att gruppen är mycket nöjda.

2 Teoretiska frågor

Besvara följande teoretiska 7 frågor:
1. Kolla på följande video: https://www.youtube.com/watch?v=X9_ISJ0YpGw&t=290s , beskrivkortfattat vad en Quantile-Quantile (QQ) plot är.

QQ plot är ett punktdiagram som illustrerar om en uppsättning datapunkter eller observationer är approximativt normalfördelade där X- och Y-axeln består av teoretisk- samt stickprovs-kvantilen. Vid ett perfekt normalfördelat dataset ligger punkterna längs en rak diagonal linje längs diagrammet.

2. Din kollega Karin frågar dig följande: ”Jag har hört att i Maskininlärning så är fokus på prediktioner medan man i statistisk regressionsanalys kan göra såväl prediktioner som statistisk inferens. Vad menas med det, kan du ge några exempel?” Vad svarar du Karin?

Jag håller med Karin och lägger till att en stor anledning med statisk regressionsanalys är att man vill förstå vad som ligger bakom resultatet. I till exempel en prediktionsmodell vill man ta reda på vilka bakomliggande faktorer som styr ett visst resultat och i vilken grad.

Ex. Om vi har en modell som predikterar huspriser kan man genom statistisk regressionsanalys förstå hur mycket (koefficienter) de ingående oberoende variablerna påverkar resultatet.

3. Vad är skillnaden på ”konfidensintervall” och ”prediktionsintervall” för predikterade värden?

Konfidensintervallet är det intervall som det predikterade värdet i snitt ligger inom för modellen givet bestämda värden i de obeorende variablerna ($x_{i...}$).

Prediktionsintervallet är det intervall det predikterade värdet ligger inom givet bestämda $x_{i...}$, för en bestämd prediktion och inkluderar på så vis osäkerheten epsilon ε.

4. Den multipla linjära regressionsmodellen kan skrivas som:
$Y = \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \ldots + \beta_p x_p + \varepsilon$.
Hur tolkas beta parametrarna?

Beta parametern β_i anger hur mycket Y förändras för varje ökning av x_i dvs den tillhörande oberoende variablen för beta parametern, givet att alla andra x-variabler är konstanta.

5. Din kollega Hassan frågar dig följande: ”Stämmer det att man i statistisk regressionsmodellering inte behöver använda träning, validering och test set om man nyttjar mått såsom BIC? Vad är logiken bakom detta?” Vad svarar du Hassan?

Det stämmer till viss del. Mått som C_p, AIC och BIC estimerar test MSE (Mean Squared error för okända data) medan Adjusted R^2 endast straffar R^2 för extra variabler. Samtliga mått kan användas för att utvärdera och jämföra modeller baserat på träningsdata men är inget definitivt mått på modellens overfit/underfit. Eftersom man i statistisk regressionsanalys vill undersöka och förstå bakomliggande faktorer i modellen baserat på befintliga data behövs inte valdering/test set, däremot ställer det större krav på att träningsdata är tillräckligt representativ gentemot ny okänd data (produktionsdata) för att kunna göra en träffsäker inferens och prediktion.

6. Förklara algoritmen nedan för ”Best subset selection”

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using using the prediction error on a validation set, C_p (AIC), BIC, or adjusted R^2. Or use the cross-validation method.

I första steget initialiseras den tomma modellen, M_0, som inte innehåller några oberoende variabler och predikterar det genomsnittliga värdet för varje observation.

I andra steget sker en iterativ process för varje önskad prediktor k = 1, 2, … p.

a) Samtliga modeller som endast innehåller *k* prediktorer tränas (fitting).
b) Bland de tränade modeller utses den bästa modelln M_k baserat på lägst RSS eller störst R^2.

I det tredje och sista steget väljs en modell ut av de bästa modellerna M_0, … , M_p utifrån antingen storleken på prediktionsfelet av ett valideringsset, C_p, AIC, BIC eller adjusted R^2.

Enkelt förklarat kan man säga att Best subset selection testar alla möjliga oberoende variabelkombinationer oavset antal variabler och väljer ut den bästa utifrån prediktionsfelet, AIC, BIC eller adjusted R^2

7. Ett citat från statistikern George Box är: “All models are wrong, some are useful.”
Förklara vad som menas med det citatet.

Alla modeller oavsett hur välbyggda och robusta de är, är en simplifierad representation av verkligheten (”All models are wrong”). Däremot är syftet med modellerna att representera tillräckligt mycket av verkligheten att man kan dra nytta av modellen (”some are useful”).

3 Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.

Ämnet är väldigt intressant men också väldigt svårt och komplicerat, men otroligt många kaninhål man lätta kan gräva ner sig i hur mycket som helst. Det går lätt åt många timmar både när man experimenterar med kod och när man läser mer inom de olika sakerna i ämnet.

Rapportskrivning är en annan utmaning. Att skriva en bra rapport är svårt, dels att skriva på ett korrekt språk men även att veta vad man ska lyfta in i sin rapport, särskilt när man skrivit mycket kod. Svårt att transferera informationen till rapporten.

Man får lägga ner timmarna som krävs för att få till en färdig bra produkt ☺. Även om det alltid går att göra bättre med mer tid och nerlagda timmar.

2. Vilket betyg du anser att du skall ha och varför.
   VG. Tycker att jag verkligen försökt redogöra för och kritiskt motivera mina beslut i projektet. Hoppas att åtminstone tillräckligt mycket av mina tankar och motiveringar märks i rapporten, även när det känts som att jag inte lyckats få in allt jag velat i rapporten.

3. Något du vill lyfta fram till Antonio?

Tack för den helt klart roligaste och mest intressanta kursen hittils. Tycker det är bra att kraven blivit högre, man pressas att göra ett ännu bättre arbete. Det känns av.