Machine Learning för smartare pendling

Examensarbete av: Johannes Sunnanväder, Lunds Tekniska Högskola

Sammanfattning

Maskininlärning har blivit en del av programvaruutveckling och är ett kraftfullt verktyg som har många tillämpningar. Det här examensarbetet har utvärderat möjligheten att använda maskininlärning i en reseapplikation för att ge resenärerna resvägsförslag som sorteras utifrån deras användarhistorik istället för resvägsförslag som sorteras utifrån till exempel avgångstid eller ankomsttid.

Genom användning av Amazons maskininlärningstjänst Amazon Machine Learning har tre maskininlärningsmodeller utvecklats och utvärderats som samverkar för att möjliggöra sortering utifrån en resenärs resvägsdata. Maskininlärningsmodellerna använder sig av olika dataset och resultatet från maskininlärningsmodellerna påverkar det totala slutresultatet med olika vikter.

För att visa hur maskininlärningsmodellerna kan användas i praktiken har en prototyp på en android-applikation utvecklats. I prototypen kan en resenär söka efter resor, se det sorterade sökresultatet och välja en resa. Prototypen använder sig av resenärens personliga maskininlärningsmodeller för att sortera resultaten utifrån resenärens användarhistorik. När en användare väljer en resa sparas resan undan i användarens personliga datafiler i Amazon S3.

Problemformulering

I detta examensarbete besvaras följande frågor:

1. Vilka parametrar finns det som kan påverka en resenärs val av resväg?
2. Vilken eller vilka transformationsrecept ska användas i modellen för att önskat resultat ska uppnås?
3. Hur ska man gå tillväga för att, utifrån data från maskininlärningsmodellen, sortera resultaten efter relevans?

Metod

Examensarbetet har använt sig av en modifierad version av den agila utvecklingsmodellen scrum som arbetsprocess. Examensarbetet har varit uppdelat i tre faser: Informationsinsamligsfas, Utvecklingsfas och Rapportskrivningsfas.

Identifiering av parametrar
Vid identifiering av parametrar hölls först en diskussion med handledaren på företaget där en grund för parametrarna togs fram. Vid diskussionen låg fokus på vilka parametrar som kan påverka en resenärs reseupplevelse. Diskussionen följdes av en analys där de parametrar som kan användas i en maskininlärningsmodell plockades ut. I analysen studerades den data som var möjlig att hämta från Google Directions API samt från utomstående källor såsom SMHI, och utifrån detta valdes de parametrar som är möjliga att hämta från datan.

Parametrarna begränsades sedan ytterligare för att kunna hanteras inom ramen för examensarbetet. För att begränsa parametrarna studerades först datan som hämtades från Google Directions API, och de parametrar som gick att hämta därifrån plockades ut. Utifrån de nya parametrarna valdes sedan de sex parametrar som examensarbetaren ansåg ha störst påverkan på en resenärs val av resa baserat på personliga resvanor.

Generering av data
För att ha data som maskininlärningsmodellerna kunde använda som träningsdata samt för att evaluera maskininlärningsmodellerna med gjordes först en sökning efter publika dataset innehållande information om resenärers resvanor. Då inget sådant dataset hittades så togs ett beslut om att generera egen data. För detta skapades ett Java-program som hämtar in information om ett antal resor och väljer sedan en utav dessa baserat på förinställda preferenser. Resdatan sparas sedan undan i en csv-fil. Syftet var att skapa data som efterliknar hur datan skulle sett ut om den var baserad på en verklig resenärs resvanor.

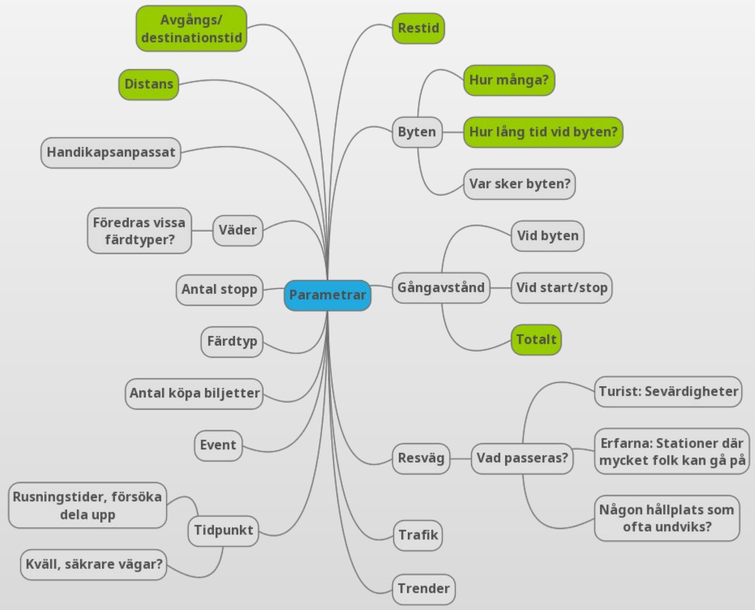
Utveckling av maskininlärningsmodeller
Baserat på den genererade datan utvecklades maskininlärningsmodeller. När en maskininlärningsmodell ska utvecklas i Amazon Machine Learning så måste ett bra transformationsrecept skapas.Transformationsreceptet bestämmer hur maskininlärningsmodellen ska transformera variablerna från indatan innan de används i maskininlärningsmodellen. Ett antal olika transformationsrecept testades och utvärderades för att slutligen hitta ett som gav ett tillfredsställande resultat.

Utveckling av prototyp av en android-applikation
Prototypen av en android-applikation utvecklades i Android Studio. I applikationen skapades funktionalitet för att söka efter resor, se sökresultaten och välja en resa. När sökresultaten presenteras går alla resor igenom maskininlärningsmodellerna i par om två resor enligt Figur 2, för att sedan sorteras utifrån deras tilldelade poäng.

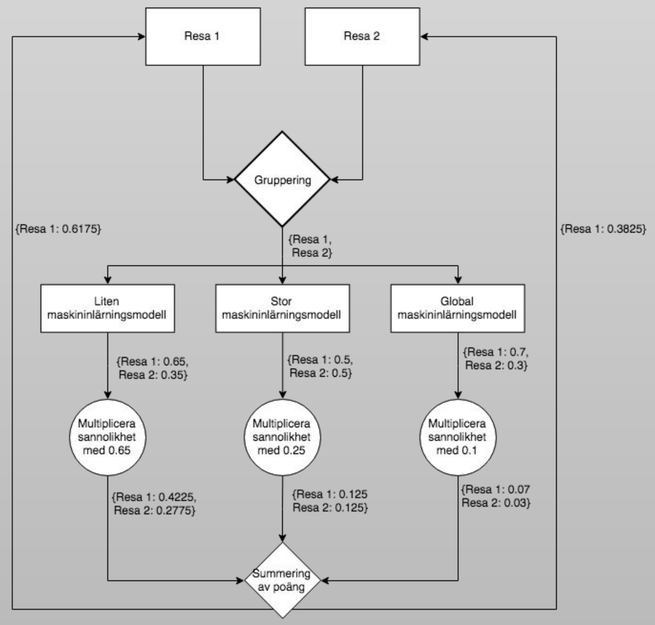
Utvärdering och tester på maskininlärningsmodellerna
För att utföra utvärderingar och tester på maskininlärningsmodellerna genererades dataset för fyra olika grupper av preferenser samt ett evalueringsset för varje grupp.

Vid utvärdering av maskininlärningsmodellerna användes den inbyggda evalueringen i Amazon Machine Learning. Olika storlekar på dataseten testades och deras precision noterades.

För att testa resultaten från maskininlärningsmodellerna användes prototypen på en android-applikation. Testet utfördes genom att söka efter resor med olika maskininlärningsmodeller aktiva och jämföra sökresultaten med de aktiva maskininlärningsmodellernas preferenser. Tester utfördes även för att se hur snabbt maskininlärningsmodellerna reagerade på förändring i preferenserna. För att testa detta kombinerades två dataset med olika preferenser, där 10 rader successivt byttes ut från det ena datasetet till det andra och ett test utfördes.



Figur 1 - Parametrar som kan påverka en resenärs val av resväg



Figur 2 - Visualisering av processen för att generera förutsägelser för resor

Resultat

Resultatet från examensarbetet är en prototyp av en android-applikation som använder sig av maskininlärningsmodellerna för att sortera sökresultaten utifrån resenärernas användarhistorik.

Parametrar som kan användas i maskininlärningsmodellen
I Figur 1 visas de parametrar som identifierats i examensarbetet som kan påverka en resenärs val av resväg. Parametrarna som är markerade med grön bakgrund är de som används i de maskininlärningsmodeller som har utvecklats i examensarbetet.

Transformationsrecept
Samtliga maskininlärningsmodeller använder sig av samma transformationsrecept. Receptet som tagits fram i examensarbetet består av tre “outputs” av parametrar som maskininlärningsmodellen använder sig av under träningsprocessen:

1. De otransformerade parametrarna.
   – Används som en utgångspunkt för att maskininlärningsmodellen ska kunna hitta mönster.
2. En sträng som består av en resas alla parametrar.
   – Används för att maskininlärningsmodellen ska kunna hitta mönster för hur en vald resas parametrar oftast ser ut.
3. Strängar som består av bindningar mellan två resor parametrar.
   – Används för att maskininlärningsmodellen ska kunna hitta mönster baserat på hur en vald resas parametrar ser ut i relation till övriga resor, till exempel färre byten.

Maskininlärningsmodeller
För en resenär kombineras tre maskininlärningsmodeller. Genom att använda tre maskininlärningsmodeller kan ett bättre resultat ges då maskininlärningsmodellerna behandlar olika typer av dataset. Maskininlärningsmodellerna är viktade för att påverka resultatet olika mycket beroende på vilken typ av data de behandlar. Följande maskininlärningsmodeller används:

- Liten maskininlärningsmodell:
  – Baseras på ett dataset som består av en resenärs 100 senaste datarader (cirka 15 resor). Syftet med den lilla maskininlärningsmodellen är att snabbt reagera på förändringar i en resenärs preferenser, vilket den gör genom att endast hantera små mängder resor. Den här maskininlärningsmodellen står för 65% av det slutgiltiga resultatet, då man vill att förändringar i preferenser snabbt ska speglas i resultatet.
- Stor maskininlärningsmodell:
  – Baseras på ett dataset som består av en resenärs alla resor. Syftet med den stora maskininlärningsmodellen är att ge en grund för vilka typer av resor en resenär föredrar, vilket den gör genom att hantera en resenärs alla resor så förutsägelser baseras på de preferenser en resenär har haft under lägst tid. Den här maskininlärningsmodellen står för 25% av det slutgiltiga resultatet, då man inte vill att tillfälliga förändringar i en resenärs preferenser ska påverka resultatet, men om förändringarna håller i sig så ska resultatet påverkas mer av den lilla maskininlärningsmodellen.
- Global maskininlärningsmodell:
  – Baseras på ett dataset som består av alla resenärers resor. Syftet med den globala maskininlärningsmodellen är att ge en överblick över hur resvanorna för alla resenärer ser ut. Den ska främst fungera som en form av utslagspoäng för två resor som anses vara lika bra matchningar enligt resenären stora och lilla maskininlärningsmodell. Den fungerar även som en maskininlärningsmodell för resenärer som inte har några sparade resor, då den globala maskininlärningsmodellen då är den enda aktiva. Den här maskininlärningsmodellen står för 10% av det slutgiltiga resultatet, då man inte vill att den ska påverka resenärernas sökresultat så att de blir felaktiga om deras preferenser inte stämmer överens med preferenserna i det globala datasetet, men att den fortfarande fungerar som en form av utslagspoäng.

Slutsats

Från examensarbetets resultat dras slutsatsen att det är fullt möjligt att använda maskininlärning för att förbättra hur resvägsförslagen presenteras i en reseapplikation. Det är då viktigt att maskininlärningsmodellerna har hög träffsäkerhet för att inte ge resenären irrelevanta resvägsförslag vilket ger en sämre användarupplevelse.

För att ge maskininlärningsmodellerna hög träffsäkerhet så behövs parametrar som påverkar en resenärs val av resväg som kan användas i maskininlärningsmodellerna. De parametrar som tagits fram i examensarbetet visas i Figur 1. Då inget dataset med verklig resdata fanns tillgängligt och parametrarna baseras på en diskussion mellan examensarbetaren och handledaren på företaget är det möjligt att det finns fler parametrar som kan användas. Det skulle varit fördelaktigt om examensarbetet skulle börjar med en datainsamling för att få riktig data att jobba med.

För att få hög träffsäkerhet i Amazon Machine Learning behövs förutom parametrar ett bra transformationsrecept. Transformationsrecept är en viktig del i Amazon Machine Learning. I examensarbetet har ett transformationsrecept tagit fram där två transformeringar utförs på parametrarna. Den första transformeringen korsar parametrarna mellan två resor för att visa mönster över vilka parametrar en resenär föredrar. Den andra transformeringen binder ihop alla parametrar för en resa till en lång sträng för att visa mönster för de enskilda resorna. Receptet som tagits fram i examensarbetet ger en bra träffsäkerhet för maskininlärningsmodellerna.

tretton37