C18 Journey to zero - Predict electricity consumption

Kristofer Klassen

Alo Martin Pallase

Michael Kevin Karlson

Baltikumi suurim ettevõte Enefit soovib aidata oma kliente elektri kasutamises ja et oleks väiksem jälg loodusele. Nimel viimasel ajal on energia hinnad läinud väga kõrgeks ja on vaja võimalusi, kuidas vähendada elektri peale minevat kulu ja kuidas vähendada ökoloogilist jalajälge. Elektri maksumust ja ökoloogilist jalajälge on võimalik vähendada, kui optimiseerida energia kasutust.

Eesmärk oleks luua energia kasutuse ennustamiseks mudel. Mida saaksid Enefiti kliendid kasutada oma majapidamise jaoks, et näha ennustust, kuna kellaajaliselt nad kõige rohkem energiat kasutavad ja selle mudeli abil muuta oma energia kasutust rohkem jätkusuutlikuks.

Eesmärk oleks saavutatud, kui klientide energia kasutus muutub nende jaoks rohkem jätkusuutlikumaks, mida hindab klient. Samuti on eesmärk saavutatud nii Enefiti kui ka Enefiti klientide jaoks, kui nende ökoloogiline jalajälg muutub väiksemaks, kasutades seda mudelit, mille abil saaks muuta majapidamise elektri kasutust.

Meie kasutada on Enefiti poolt kolm andmefaili. Üks andmefail treenimiseks, teine kus ennustada ja kolmas on esitamiseks, kus on kellaaja kõrval elektri kasutuse arv. Samuti on eesmärgi saavutamiseks Kristofer, Alo ja Michael.

Eesmärk peab olema saavutatud 9 detsembriks 2022. Reeglistiku pole, see on avatud koodiga ehk kõigil on õigus ja luba näha saavutatud koodi ja mudelit. Et eesmärk oleks saavutatud peab tehtud koodi abil tulema korrektne mudel, mis ennustab ette antud seitsme päeva energia kasutust kellaajalise täpsusega.

Riskid oleks elektri kadumine kodus. Selle lahendus oleks kirjutada koodi kellegi teise juures või Deltas. Samuti oleks probleemiks, kui peaks internet kaduma, siis oleks samad

lahendused, mis elektri kadumisel. Kolmandaks probleemiks võib tulla aja puudu jääk, et ülesanne õigeks ajaks valmis saada. Selle lahendus oleks teha korralik ülesannete jagamine, et ei tekiks seda probleemi.

Kasutusel olev terminoloogia inglise keeles:

- time - definition of example_id
- temp - Air Temperature (°C)
- dwpt - The dew point in °C
- rhum - The relative humidity in percent (%)
- prcp - The one hour precipitation total in mm
- snow - The snow depth in mm
- wdir - The wind direction in degrees (°)
- wspd - The average wind speed in km/h
- wpgt - The peak wind gust in km/h
- pres - The sea-level air pressure in hPa
- coco - The weather condition code
- el_price - the electricity price in Estonia on that hour (€/kWh)
- consumption - the electricity consumption (kWh)

Sellel projektil rahalisi kulusid ja tulusid pole.

Andmete töötlemise tulemusel peaks tekkima ennustus mudel, mis näitab ennustatavat energia kasutust ühe majapidamise kohta. Samuti on vaja teha projekti poster, mida esitleda Sissejuhatus andmeteadustesse raames. Töötlemiseks on kaks andmestikku mis sisaldavad ilma, elektri hindu ja elektri kasutust perioodil 2021-09-01 00:00 - 2022-08-24 23:00. Ennustamise mudeli jaoks on kasutada periood 2022-08-25 00:00 - 2022-08-31 23:00, kus on antud elektri hind ja ilma andmed, kuid puudub kasutuse andmed. Ning ülesandeks on ennustada järgneva seitsme päeva energia kasutus selle majapidamise raames.

Andmestiku koostamine on edukas, kui tehtud mudel on võimalikult väikse erinevusega ennustava kasutusega versus päris kasutus kogusega. Antud projekti puhul kasutatakse „**Mean Absolute Error - MAE**“**.** Mudeli edukust ja täpsust hindab Enefit.

Andmete mõistmine

Meie kasutada on ülesande koostaja poolt etteantud puhastamata andmed, mis asuvad failis *train.csv.* Fail on Exceli formaadis, andmed on kättesaadavad ja loetavad ning meie valitud andmekaeve keskkond loeb formaadist andmeid korrektselt sisse.

Andmeid on 8592 rida, mis on tunniajase intervalliga esitatud. Igal real on kuupäev koos kellaajaga ja tol hetkel mõõdetud õhutemperatuur, kastepunkt, õhuniiskus, sademed millimeetrites, lume sügavus, tuule suund, keskmine tuule kiirus, suurim tuule puhangu kiirus, õhurõhk meretaseme kõrguselt, ilma olukorra kood, elektrihind sellel tunnil Eestis, elektrikulu kWh kohta.

Peamised andmed, mida meie enda projekti eesmärgi saavutamiseks kasutame on temperatuur, õhuniiskus, tuulesuund, tuulekiirus, elektrihind ja elektrikasutus. Andmete hulgas on ka atribuute mille väärtuseid ei saa või ei ole mõtet kasutada kuna puudub piisav info nende kohta. Meie eesmärgi saavutamisel ei anna sademed piisavalt kasulikku infot. Samuti ei leia suurt kasutust ka ilma olukorra kood kuna ei anna lisainfot teiste atribuutide kõrvalt.

Andmed vajavad kindlasti puhastamist. Elektrihindade osas leidsime minimaalseks elektri hinnaks 7.000000000000001e-05, mis ei ole realistlik hind elektrile. Suurim elektri hind oli 4.0, mis vastab tõele antud ajavahemiku puhul. Aasta keskmine temperatuur antud andmetes vastab ootustele seega võime eeldada, et temperatuuri andmed on korrektsed. Leidsime ka elektri kasutuse kogu perioodi vältel, milleks on 358 päeva ning saime kasutusele kokku 8988 kWh, mis on samuti keskmise majapidamise aastane energia tarbimine seega antud andmete põhjal tehtav mudel peaks teoorias sobima mitte ainult konkreetsele majapidamisele, millelt andmed kogutud on vaid ka teistele sarnastele elamispindadele. Sademete hulgast leidsime 116 positiivset väärtust ja 2043 ’0’ väärtust 8592 kirje hulgast, mis ei anna meile väga palju lisainfot eesmärgi saavutamisel.

Kokkuvõtvalt on meie kasutusel antud andmed korrektsed, vajavad lihtsalt puhastamist. Saame antud andmeid enda eesmärgi saavutamiseks kasutada ning lisaandmeid ei vaja.

Projekti planeerimine

1. Task - andmete kogumine. Meie projekti andmed pärinevad kaggle'ist ning selle kogumisele ei kulu palju aega.
2. Task - Andmete puhastamine. Kuna projekti andmed pärinevad ilma mõõtmistulemustest, võib olla puuduvad andmed mingitel kellaaegadel või ei ole andmed sobivas vormingus, et mudel neist õppida saaks. Selle jaoks tuleb kasutada Pandas-t, et viia andmed sellisele kujule, et need oleksid mudeli jaoks loetavad. Lisaks võib andmete järjekorra randomiseerida, et mudel ei looks valesi arvamusi andmete järjestusest. Lõpuks tuleks andmed jagada kaheks osaks, treening-osa ja testimis-osa, et olek nn *baseline*, mille vastu võrrelda mudelit. Selle jaoks võiks kuluda ~1 tund 1 team member.
3. Task - Mudeli valimine ja treenimine. Tõenäoliselt tuleb meil treenida mitut mudelit ja katsetada nende efektiivsust, et leida parimad mudelid. Mudelid võtame sklearn raamatukogust (*library*). ~3 hours each team member.
4. Task - parameetrite tuunimine. Kui on leitud parimad mudelit meie andmete jaoks, siis tuleks nende mudelite parameetreid tuunida, et suurendada nende korrektsust. Peale tuunimist, valime kõige parema mudeli ning kasutame seda lõplike eelduste tegemiseks. ~2 hours each team member
5. Task - Eelduste tegemine. Kui on leitud parim mudel siis tuleb kasutada seda, et teha eeldused võistluse jaoks. ~1 hour 1 team member
6. Task - Tulemuse esitamine. Kui on olemas sobiv mudel ja selle tehtud eeldused siis viimase ülesandena peame esitama need kaggle võistlusesse. ~<1 hour 1 team member.