CODA

Predictive Models with Interpretability and COncept Drift Analytics

Publik rapport

Författare: Jesper Brauer, Erik Frisk, Tony Lindgren, Anders Vesterberg, Pär Sundbäck
Datum: 2020-03-29
Projekt inom *Effektiva och uppkopplade transportsystem*

FFI Fordonsstrategisk Forskning och Innovation
VINNOVA Energimyndigheten TRAFIKVERKET
FKG VOLVO SCANIA VOLVO

Innehållsförteckning

Kort om FFI

FFI är ett samarbete mellan staten och fordonsindustrin om att gemensamt finansiera forsknings- och innovationsaktviteter med fokus på områdena Klimat & Miljö samt Trafiksäkerhet. Satsningen innebär verksamhet för ca 1 miljard kr per år varav de offentliga medlen utgör drygt 400 Mkr.

För närvarande finns fem delprogram; Energi & Miljö, Trafiksäkerhet och automatiserade fordon, Elektronik, mjukvara och kommunikation, Hållbar produktion och Effektiva och uppkopplade transportsystem. Läs mer på www.vinnova.se/ffi.

1 Sammanfattning

Projektets parter var Scania CV, Linköpings universitet (Institutionen för systemteknik), Stockholms universitet (Institutionen för datavetenskap) och Kungliga tekniska högskolan (Skolan för informations- och kommunikationsteknik). Projektet pågick mellan november 2017 till och med februari 2021 med en total budget på 18,2 MSEK.

Syftet med projektet var att ta fram mer kunskap runt användningen av prediktiva modeller för fordon med fokus på datadrift och tolkningsbarhet av modeller. En annan viktig fråga för projektet har varit processer för hantering av modeller och data.

Projektet var uppbyggt runt sex arbetspaket med följande innehåll och resultat:

1. **Projektstyrning** – Arbetspaketets huvudsyfte var att operativt leda projektet, samt ansvara för leveranser till Vinnova.
2. **Datadrift och dynamisk data** – Grundfrågeställningen som studerats i det här arbetspaketet rör hur man utvecklar data-drivna prognosticerande modeller baserat på data i ständig förändring, här kallat dynamisk data. Förändringen i data kommer av att nya data ständigt strömmas in och registreras, i det här fallet då varje lastbil sänder användnings- och verkstads-data kontinuerligt. Resultaten visar att det finns prediktiv information i sekvenser av utläsningar och att det är möjligt att förbättra prediktiv prestanda. Resultaten visar också att vinsten inte är så stor som man kan vänta sig, och det beror sannolikt på att komponentens livslängd främst påverkas av den totala belastningen och inte på transienta förlopp. Arbetspaketet har också studerat reducering av data utan att tappa information vilket ger mer kondenserade data som är enklare att bygga modeller av samt ger ökad möjlighet till förståelse av modellen.
3. **Tolkningsbarhet av modeller** – Arbetspaketet har jobbat med framtagning av metoder och kunskap för att främja tolkningsbarheten av datadrivna prediktionsmodeller. Här har en del av arbetet ägnats åt hur prediktionsmodeller skall kunna nyttja data som kommer trunkerat i ett histogramformat. Arbetspaketet har också undersökt och skapat kunskap om hur prediktionsmodeller med histogramdata kan presenteras för användare, detta inkluderar en programvara med olika visualiserings vyer för att åskådliggöra t.ex. varför en viss prediktion görs av en modell i en specifik beslutsfunktion.
4. **Driftsättning av prognostiska modeller** – Arbetspaketet har tagit fram ett generaliserat ramverk/arkitektur för driftsättning av modeller. AWS och maskininlärningsplattformen SageMaker valdes som bas för att utforska driftsättning och tillhörande processer eftersom AWS-plattformen bedömdes ha mycket stor framtidspotential. Med AWS SageMaker kan man bygga, träna och produktionssätta maskininlärningsmodeller i säker och skalbar produktionsmiljö. Som reellt användningsfall valdes prediktering av livslängd för generatorn i Scanias lastbilar och bussar. En arkitektur med pipeline för hämtning av data, rensning och kontroll, träning av modell och produktionssättning i form av webbtjänst implementerades med byggstenar från AWS. Arbetspaketet har använt sig av metoder som fostrar till modulär design och leder till kvalitet, liknade de som används vid vanlig programutveckling.
5. **Effektiviserad datainsamling** – Huvudleveransen från detta arbetspaket var en demonstrator för datainsamling på ett konfigurerbart sätt. Datat kunde sedan användas för att upptäcka avvikelser ombord eller som bas för andra användningsfall, som prediktivt underhåll etc. Demonstratorn har använts för att skaffa kunskap om utvärdering av principer för kravställningen, datadefinitioner och arkitektur för datainsamling. Demonstratorn innehåller arkitektur och implementation av en ’edge’-nod kopplad till fordonets CAN nätverk som kan samla in och spara godtyckligt CAN-data och leverera det som tidsserier eller histogram till en molntjänst. Molntjänsten tar emot

data och utvärderar det samt bygger en modell för detta. Modellen kan sedan laddas ner till fordonet och där utvärdera data i farten.

6. **Data warehouse** – Arbetspaketet arbetade med hur externa användare ska få access och tillgång till ett företags data. Som användningsfall användes en extern användare inom projektet och arbetet resulterade i processer och guidelines som kan användas för liknande situationer i framtiden.

De väsentliga projektmålen har uppfyllts och därigenom har projektet avsevärt stärkt forskningsområdet inom maskininlärning och prognostisk – och då särskilt inom konceptdrift och tolkningsbarhet hos modeller. Projektet levererade två doktorsavhandlingar, runt 15 vetenskapliga artiklar, och åtta examensarbeten. Dessutom tog projektet fram programvara för tolkningsbarhet, ramverk/arkitektur för driftsättning av maskininlärningsmodeller, en demonstrator för datainsamling och processer/guidelines för datadelning med externa användare. Totalt hölls 17 workshops, varav ett var publikt, för kunskapsöverföring och spridning av resultaten.

2 Executive summary in English

The parties in the project were Scania CV, Linköping University (Department of Systems Engineering), Stockholm University (Department of Computer Science) and Royal Institute of Technology (School of Information and Communication Technology). The project ran from November 2017 until February 2021, with a total budget of 18.2 million SEK.

Data-driven methods based on machine learning have shown their strength in many applications. In a previous research project, IRIS funded by FFI and Scania, it has been demonstrated how data-driven methods could be used to plan optimal maintenance.

This project has highlighted two characteristics that are important and need further research, the ability to handle concept drift in data and the possibilities for interpretability. Concept drift arises when the fleet of products that the models have been created to predict the health of changes over time. This occurs for example naturally in the automotive industry when new modules are available for sale or the launch of updated features. Strongly linked to concept drift is interpretability. In order for engineers to understand if their design changes affect the models, it would be a strength if they could understand how different variables and properties affect the models' outcomes.

The project comprised six work packages with the following content and results:

1. **Project management** – The main purpose of the work package was to operationally lead the project, as well as be responsible for deliveries to Vinnova.
2. **Data drift and dynamic data** – The basic issue studied in this work package concerns how to develop data-driven predictive models based on data in constant change, here called dynamic data. The change in data comes from the fact that new data is constantly streamed in and registered, in this case when each truck is sending operational and workshop data continuously. The results show that there is predictive information in sequences of readings and that it is possible to improve predictive performance. The results also show that the gain is not as large as might be expected, and this is probably due to the fact that the component's service life is mainly affected by the total load and not by transient events. The work package has also studied the reduction of data without losing information, which provides more condensed data. This data is easier to build models from and provides an increased opportunity for understanding the model.
3. **Interpretability of models** – The work package has worked with the development of methods and knowledge to facilitate the interpretability of data-driven predictive models. Part of the work has been devoted to how predictive models should be able to use data

that comes truncated in a histogram format. The work package has also investigated and created knowledge about how predictive models with histogram data can be presented to users, this includes software with different visualization views to illustrate e.g. why a certain prediction is made by a model in a specific decision-making situation.

4. **Deployment of prognostic models** – The work package has developed a generalised framework/architecture for deploying models. AWS and the machine learning platform AWS SageMaker were chosen as a base to explore deployment and associated processes, since the AWS platform was considered to have a great potential for the future. With AWS SageMaker, you can build, train and set up machine learning models in a secure and scalable production environment. Prediction of service life for the generator in Scania's trucks and buses was chosen as a use case. An architecture with a pipeline for data collection, cleaning and control, model training and production set-up in the form of web service, was implemented with building blocks from AWS. The work package has used methods that foster modular design and lead to quality, similar to those used in software development.
5. **Configurable data collection and anomaly detection** – The main deliverable from this work-packages was a demonstrator, showing how to collect data in a configurable way. This data can then be used for anomaly detection onboard, or as a base for other use-cases, like predictive maintenance. The demonstrator has been used to acquire knowledge about evaluation of principles for setting requirements, data definitions and architecture for data collection. The demonstrator contains architecture and implementation of an ‘edge’ node connected to the vehicle's CAN network that can collect and save any CAN data, and deliver it as time series or histograms to a cloud service. The cloud service receives data and evaluates it, and builds a model for it as well. The model can then be downloaded to the vehicle and evaluate data in real time.
6. **Data warehouse** – The work package has investigated on how external users should get access to a company's data. An external user within the project was used as a use case, and the work resulted in processes and guidelines that can be used for similar situations in the future.

The main project goals have been met, and thereby the project significantly has strengthened the field of research in machine learning and prognostics, and especially in the field of concept drift and interpretability of models. The project has delivered two doctoral thesis, around 15 scientific articles, and eight degree thesis. In addition, the project developed software for interpretability, framework/architecture for deploying machine learning models, a demonstrator for data collection and processes/guidelines for data sharing with external users. A total of 17 workshops, one of which was public, were held for knowledge transfer and dissemination of the results.

3 Bakgrund

För att effektivisera transport av gods och människor krävs att fordon kan utnyttjas enligt plan samt att de inte blir stillastående på väg, med risk för följdolyckor och köbildning. Om ett fordons hälsostatus kan förutspås eller prognostiseras under en lämplig underhållshorisont finns potential för att effektivisera underhåll, öka fordonens tillgänglighet, och minska risker för dyrbara fordonshaverier och förstörd last. Datadrivna modeller är ett resurseffektivt sätt för att ta fram prognoser för kritiska komponenters (t.ex. startbatteriet) hälsostatus i ett givet fordon med specifik användningshistorik och konfiguration.

Eftersom relativt få specifika mätningar sparas för varje enskilt fordon används data från hela fordonsflottan för att ge tillförlitliga resultat. Ett problem med datadrivna modeller är att de ger prediktioner utan explicit förklaring. Därför skulle tilliten och förståelsen kunna ökas avsevärt om resultaten är tolkningsbara, dvs. att det presenteras begripliga förklaringar till vad som ligger till

grund för prediktionerna, t.ex. varför batterihälsan hos ett specifikt batteri bedöms vara bra eller dålig.

En annan utmaning när system för prediktivt underhåll är driftsatta är att de ska fungera kontinuerligt år efter år trots att nya komponenter och system införs på de sålda produkterna, eller andra komponenter tas bort eller modifieras, och loggade data förändras över tid, så kallat datadrift. Det är därför viktigt att de prognostiska algoritmerna kan hantera dessa förändringar på ett effektivt sätt. Till exempel, hur uppdateras modeller på ett beräkningseffektivt sätt när ny data strömmar in? Om nya komponenter införs, hur bra kan dessas hälsa prognostiseras? Om variabler slutar loggas på nya fordon, hur bevaras den kunskap som fanns i gammal data?

I och med att analys av data blir viktigare för framtida utveckling av tjänster och funktioner är det för svensk industri vitalt att kunna samverka med starka akademiska forskningskluster. För att kunna genomföra detta krävs att data och beräkningskapacitet kan upplåtas till parter utanför industrin samtidigt som data skyddas både ur ett kommersiellt perspektiv och för uppfyllande av personlig data lagkrav.

4 Syfte, forskningsfrågor och metod

Syftet med projektet var att ta fram mer kunskap runt användningen av datadrivna prediktiva modeller för fordon. Projektet har huvudsakligen fokuserat på två forskningsfrågor:

- Hur ska datadrift hanteras? Datadrift uppkommer då flottan av produkter som modeller är skapade för att prediktera hälsan hos ändras över tid, t.ex. uppkommer detta naturligt inom fordonsbranschen där nya moduler tillgängliggörs för försäljning eller uppdaterade funktioner lanseras.
- Hur kan resultaten göras tolkningsbara? Tolkningsbarheten är starkt kopplat till datadrift. För att ingenjörer ska kunna förstå om deras designändringar påverkar modellerna vore det en styrka om de kan förstå hur olika variabler och egenskaper påverkar modellernas utfall.

En annan viktig fråga för projektet har varit processer för hantering av modeller och data.

Beskrivningar av metoderna som har använts i projektet presenteras under de olika arbetspaketen i kapitel 5 och 6.

5 Mål

Det övergripande målet var att stärka forskningsfältet inom maskininlärning och prognostik och specifikt inom området datadrift och tolkningsbarhet av modeller med fokus på underhåll för tunga fordon. Projektet var uppbyggt runt sex arbetspaket med följande mål:

1. **Projektstyrning** – Operativ ledning av projektet och ansvar för leveranser till Vinnova inklusive rapporter och presentationer.
   Leveranser: a) Administrativ ledning av projektet, b) Anordnande av extern workshop, c) Rapportering till Vinnova, d) Kallande till ledningsgruppsmöte
2. **Datadrift och dynamisk data** – Framtagning av metoder och kunskap för analys av datadrift och dynamisk data. Metoderna ska testas på riktiga data för en flotta av tunga fordon. Utredning av vilka kommersiella verktyg som kan användas för analys samt identifikation av de krav som existerar för att datadrift ska kunna hanteras över en tid som sträcker sig över minst 10 år.
   Leveranser: a) Vetenskapliga artiklar, b) Tre examensarbeten, c) Demonstrationer av metoder, d) Genomförande av workshops
3. **Tolkningsbarhet av modeller** – Framtagning av metoder och kunskap för analys och förbättring av tolkningsbarhet. Metoderna kommer att utvärderas med avseende på ett

antal modeller framtagna vid Scania samt inom IRIS-projektet.
Leveranser: a) Vetenskapliga artiklar, b) Tre examensarbeten, c) Demonstrationer av metoder, d) Genomförande av workshops

4. **Driftsättning av prognostiska modeller** – Framtagning och driftsättning av en arkitektur baserat på PMML-modeller (Predictive Model Markup Language). Förutom PMML ska även en analys av alternativa standarder genomföras. Kraven på arkitekturen är att den ska kunna hantera olika typer av PMML-modeller, så som random forest och neurala nätverk, samt förbehandling och efterbehandling av data. Vidare ska processer tas fram för att effektivt kunna driftsätta nya modeller med upprätthållande av kvalitet och till exempel ska det var möjligt att driftsätta modellerna för delar av en flotta. Arkitekturen ska även möjliggöra kontrollerade experiment eller så kallade A/B-tester. Simulering av utfall är också ett krav för att innan driftsättning kunna verifiera utfallet. I arkitekturen ingår både insamlingen av data från fordon, evaluering av modellerna, beräkning av underhållsschema och tillhörande analys av datadrift utvecklad inom de andra arbetspaketen.
   Leveranser: a) Driftsättning av generaliserat ramverk för att evaluera och utvärdera modeller, b) Test av hur analys av datadrift ska kunna implementeras, c) Identifiering av krav
5. **Effektiviserad datainsamling** – Framtagning av en process som kan resultera i bättre kravställning. Vidare ska riktlinjer tas fram för hur data ska samlas in och vilken typ av data som ska samlas in, med särskilt beaktande av hur ny data effektivt kan tillföras redan producerade produkter. Ett resultat kommer vara en arkitektur för flexibel datainsamling. Dessutom ska möjligheterna att skapa artificiella utfall genom till exempel simulering eller experiment utredas.
   Leveranser: a) Identifikation av krav b) Tre examensarbeten, c) Rapport beskrivande de behov som krävs av datautläsning d) Implementation av demonstrator för flexibel, dynamisk hantering av data från fordon. Demonstratorn ska innehålla mekatroniskt system on-board, kommunikationsenhet, telekommunikation och backoffice-system
6. **Data warehouse** – Data warehouse för data analytics
   Leveranser: a) Uppsättning av säker anslutning för externa parter till industripartens datasjö och beräkningskluster, b) Utveckling av relevant behörighet för externa partners, c) Implementation av system uppfyllande allmänna dataskyddsförordningen och andra krav

6 Resultat och måluppfyllelse

I detta avsnitt presenteras sammanfattade resultat av varje arbetspaket samt en lista på leveranserna.

Arbetspaket 1 – Projektstyrning

Arbetspaketets huvudsyfte var att operativt leda projektet, samt ansvara för leveranser till Vinnova inklusive rapporter och presentationer. Totalt hölls 13 workshops med projektgruppen, samt ett flertal kortare pulsmöten. Varje workshop inleddes först med ett ledningsgruppsmöte och därefter presenterades delresultat och aktuella frågeställningar från varje arbetspaket. Mötena var fysiska ända tills mars 2020 då COVID-19 omöjliggjorde större fysiska träffar. På grund av COVID-19 (och dess konsekvenser på arbetet på Scania) förlängdes projektet från 2020-10-31 till 2021-02-28. Scania gjorde också en mindre budgetförändring på grund av konsekvenserna av pandemin. Hela projektet avslutades med två workshops:

- Publik workshop 2021-03-12 med presentationer av projektresultaten, presentation från Stena Line's AI-chef och en paneldiskussion med experter inom digitalisering, data transformation och prediktiv modellering.
- Intern workshop på Scania 2021-03-19 med presentationer av projektresultaten (på en mer detaljerad nivå än vid den publika workshopen).

Arbetspaket 2 – Datadrift och dynamisk data

Grundfrågeställningen som studerats i det här arbetspaketet rör hur man utvecklar data-drivna prognosticerande modeller baserat på data i ständig förändring, här kallat *dynamisk data*. Förändringen i data kommer av att nya data ständigt strömmas in och registreras, i det här fallet då varje lastbil sänder användnings- och verkstads-data kontinuerligt. Detta är en trend som kan förväntas fortsätta då digitaliseringen av transportsektorn ständigt utvecklas, och mer och mer data samlas in. Med dynamisk inkommande nya data, *strömmade data*, kommer karaktären på data förändras över tid, exempelvis när nya fordonstyper tas i bruk och ska prognostiseras baserat på data från andra fordonstyper.

Man kan grovt särskilja på två typer av datadrift, plötsliga förändringar och gradvisa förändringar. Plötsliga förändringar kan motsvara introduktionen av en ny lastbilsmodell som har en annan struktur på data. Gradvis förändring kan ske exempelvis med långsamt förändrad komposition av fordonsflottan som genererar data. Speciellt viktigt är det att hantera obalanserade data, dvs. när det finns mycket fler datapunkter för fullt fungerande komponenter jämfört med komponenter som behöver åtgärdas. Detta är typiskt i fallet med prognostik av hälsostatus där fullt fungerande fordon är i stor majoritet för insamlade data.

En första huvudfråga är därför hanteringen av strömmad data. Den information som samlas in från lastbilarna är här, och generellt, ofta av kumulativ karaktär vilket innebär att det är intressant att undersöka hur mycket mer information man får ut från sekvenser av datapunkter relativt den information som finns insamlad vid den senaste utläsningen. Detta är speciellt intressant inom fordonsapplikationen då vi ej kan förvänta oss täta utläsningar. Resultat visar att det finns prediktiv information i sekvenser av utläsningar och att det är möjligt att förbättra prediktiv prestanda vilket är visat genom att jämföra resultaten från publikationerna [1] och [4] med resultaten i [3]. Detta är väntat, men resultaten visar också att vinsten inte är så stor som man kan vänta sig, och det beror sannolikt på att komponentens livslängd främst påverkas av den totala belastningen och inte på transienta förlopp.

En viktig fråga för insamlade fordonsdata av den här typen, som ofta inte är anpassad för prognostiska modeller av den här modellen utan är av mer generisk karaktär, är vilka mätvariabler som är av intresse för prognostik av en specifik komponent. En reducering av data utan att tappa information ger mer kondenserade data som är enklare att bygga modeller av samt ger ökad möjlighet till förståelse av modellen. Detta har studerats i publikationerna [2] och [5].

En workshop om arbetspaket 2 genomfördes den 18 april 2018. På denna workshop diskuterades forskningsområdet *datadrift och dynamiska data*. Totalt deltog mer än tio personer från de olika delarna av projektet i denna workshop. Resultatet av diskussionerna låg sedan till grund för den fortsatta forskningsinriktningen i arbetspaketet under resten av projektet.

Sergii Voronovs doktorsavhandling [7], som han försvarade fredagen den 6 mars, 2020, innefattar och sammanfattar väl resultaten inom arbetspaketet. Avhandlingen består av en introduktion och inkludering av 6 publicerade arbeten. Opponent vid försvaret var Prof. Kai Goebel, PARC, USA.

Arbetspaketets bidrag till FFI:s mål

I arbetspaketet har det utarbetats metoder för att automatisera underhåll av nödvändiga komponenter för att uppfylla transportuppdrag. Metoderna är möjliga i och med den digitaliseringsprocess som pågår inom fordonsindustri, där mer och mer data görs tillgängligt för att utveckla tjänster och produkter. Specifikt har underhåll av batterier undersökts med mål att specialanpassa underhåll för varje fordon för att maximera tillgängligheten av fordonet samt maximera utnyttjandet av batteriets kapacitet. Dessa delar kopplar direkt till FFI-målen *automatisering, digitalisering* samt *elektrifiering*. Kommersialisering, utveckling av nya tjänster

och affärsmodeller baserat på insamlade data är beroende på utveckling av funktionalitet som studerats i det här arbetspaketet. Därmed är resultaten också indirekt kopplade till framtida *affärsmodeller*. De insamlade data som används i arbetspaketet för att utveckla prediktiva modeller rymmer information från en stor mängd olika typer av fordon samt användningsprofiler. Att hantera detta har varit en viktig del i arbetspaketet, speciellt med avseende på datadrift, och metoderna har visat sig användbara även för detta vilket gör resultaten relevanta för FFI-målet *anpassade fordonskoncept*.

Totalt utlovades att följande leveranser skulle ske från arbetspaketet:

- Vetenskapliga artiklar.
- Tre examensarbeten.
- Demonstrationer av metoder.
- Genomförande av workshop.

Alla dessa leveranser har blivit genomförda enligt plan förutom examensarbeten som vi inte fullt ut lyckats få till. Se avsnitt 7.2 för en sammanfattning av alla publikationer i arbetspaketet.

Arbetspaket 3 – Tolkningsbarhet av modeller

Arbetspaketet har jobbat med framtagning av metoder och kunskap för att främja tolkningsbarheten av datadrivna prediktionsmodeller. Här har en del av arbetet ägnats åt hur prediktionsmodeller skall kunna nyttja data som kommer trunkerat i ett histogramformat. Denna typ av data är väldigt vanlig vid industriella resurser som inte har hög minneskapacitet eller möjlighet att överföra stora datamängder till externa beräkningsresurser som en molntjänst.

Vidare har sedan arbetspaketet undersökt och skapat kunskap om hur prediktionsmodeller med histogramdata kan presenteras för användare, detta inkluderar en programvara med olika visualiserings vyer för att åskådliggöra t.ex. varför en viss prediktion görs av en modell i en specifik beslutsfunktion. Visualiseringen beskriver även den ”globala” modellens beteende, genom att visa vilka variabler som är viktigast när den gör ett beslut mm. Mjukvaran för tolkningsbarhet har delats med den industriella partnern i projektet.

En workshop om arbetspaketet tre genomfördes den 23:e Maj 2018, på denna workshop diskuterades forskningsområdet *tolkningsbarhet av modeller*. Totalt deltog tio personer från de olika delarna av projektet i denna workshop. Resultatet av diskussionerna låg sedan till grund för den fortsatta forskningsinriktningen i arbetspaketet under resten av projektet.

Arbetspaketets bidrag till FFI:s mål

Höja den tekniska mognadsgraden. I arbetspaketet så har det utvecklats effektiva metoder att ta fram prediktionsmodeller baserat på histogramdata, detta bidrar till FFI:s övergripande mål att höja den tekniska mognadsgraden inom området.

Människan i systemet. Arbetspaketet har bidragit med analys och kunskap om hur människan i form av underhållsexperter och utvecklingsingenjörer ska kunna jobba med prediktionsmodeller som tas fram av dataanalytiker. Mjukvaran som tagits fram för att möjliggöra förståelse för dess prediktionsmodeller har använts i ett experiment för att tillsammans med underhållsexperter och utvecklingsingenjörer för att utvärdera hur väl verktyget fungerar och hur verktyget kan förbättras, i artikeln [15] återfinns resultatet av denna undersökning. Slutligen så överlämnades denna mjukvara till Scania för att säkerställa kunskapsöverföring från akademin till industrin. En demonstration av mjukvaran skedde vid Innovation Day på Scania den 18 september 2019.

Alla vetenskapliga artiklar och examensarbeten producerade inom ramen för arbetspaketet, bidrar till att uppfylla de bägge nämnda målen ovan.

Stora delar av kunskapsbidraget som arbetspaketet bidragit med sammanfattas väl i Ram Gurungs doktorsavhandling, som han försvarade, fredagen 20 mars 2020. Vid disputationen fanns åhörare från akademin och industrin, och tack vare opponentens föredömliga arbete och Rams bidrag, så ledde diskussionerna antagligen till att flera deltagare fick nya kunskaper och möjligtvis nya idéer för framtida forskning inom området.

Totalt utlovades att följande leveranser skulle ske från arbetspaketet:

- Vetenskapliga artiklar.
- Tre examensarbeten.
- Demonstrationer av metoder.
- Genomförande av workshop.

Alla dessa leveranser har blivit genomförda enligt plan, i kapitel 7.2 vid sektion ”Arbetspaket 3 – Tolkningsbarhet av modeller” följer en lista över vetenskapliga artiklar och examensarbeten som skapats inom ramen för arbetspaketet.

Arbetspaket 4 – Driftsättning av prognostiska modeller

I den ursprungliga projektbeskrivningen beskrevs PMML som en tänkbar och viktig arkitektur för driftsättning. Efter en studie visade det sig dock att stödet för olika modeller i PMML var begränsat, i synnerhet för överlevnadsmodeller. Istället valdes AWS och maskininlärningsplattformen SageMaker som bas för att utforska driftsättning och tillhörande processer. AWS-plattformen bedömdes ha mycket större framtidspotential. Med AWS SageMaker kan man bygga, träna och produktionssätta maskininlärningsmodeller i säker och skalbar produktionsmiljö. SageMaker har många inbyggda algoritmer och hanterar versionering av modeller och A/B-test.

Som reellt användningsfall valdes prediktering av livslängd för generatorn i Scanias lastbilar och bussar. En arkitektur med pipeline för hämtning av data, rensning och kontroll, träning av modell och produktionssättning i form av webbtjänst implementerades med byggstenar från AWS. Möjligheten att använda egenutvecklade algoritmer utforskades, bland annat implementerades i SageMaker en prediktionsalgoritm av type Conformal Prediction. Denna algoritm utvecklades under projektets gång under ett annat arbetspaket.

En utvecklingsprocess har använts och beskrivits som sträcker sig från utforskande i enkel, men interaktiv notebookmiljö, till utveckling i integrerad utvecklingsmiljö. Det finns många fördelar med att tidigt använda sig av struktur och automation som stöds av utvecklingsmiljöer under utvecklingsprocessen. Projektet har även använt sig av kontinuerlig integration och produktionssättning (CI/CD). Dessa angreppssätt, precis som vid vanlig programutveckling, fostrar till modulär design och leder till kvalitet. Lokal felsökning och test av molnfunktionalitet har utretts, använts och dokumenterats.

Målet att leverera ett generaliserat ramverk/arkitektur för driftsättning av modeller har uppfyllts. Diskussioner angående test för datadrift har hållits, men implementation har inte skett. Fokus förflyttades istället mot hur underhållsschema kan skapas utifrån överlevnadsmodeller och diskussionerna om detta har gett värdefulla insikter.

Arbetspaketets bidrag till FFI:s mål

I arbetspaketet har det tagits fram ett generaliserat ramverk/arkitektur för driftsättning av maskininlärningsmodeller med fokus på hur underhållsschema kan skapas utifrån överlevnadsmodeller. Därmed finns det en direkt koppling till *Fordons- och mobilitetstjänster*.

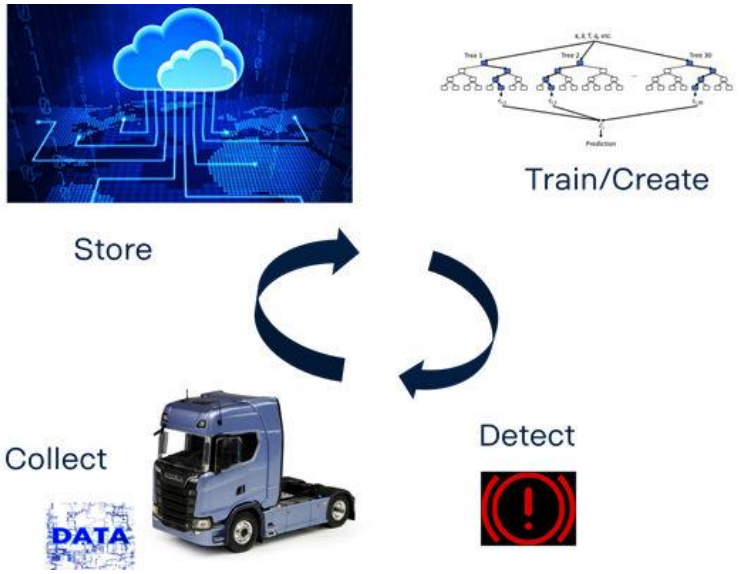
Arbetspaket 5 – Effektiviserad datainsamling

Under projektets gång har mycket ändrats både i Scania organisationen och med Scanias produkter och därför har även arbetspaketets innehåll ändrats något. Vi har haft mer fokus på

demonstratorn och genom att implementera den lärt oss och utvärderat principer för kravställningen, datadefinitioner och arkitektur för datainsamling. Även kraven för datautläsning har definierats genom att bygga demonstratorn.

Demonstratorn har också utökats lite och getts förmåga till anomalidetektion ombord med hjälp av maskinlärning i molnet.

Leveranser
Arkitektur och implementation av en ’edge’-nod kopplad till fordonets CAN nätverk som kan samla in och spara godtyckligt CAN-data och leverera det som tidsserier eller histogram till en molntjänst. Molntjänsten tar emot data och utvärderar det samt bygger en modell för detta.



Modellen kan sedan laddas ner till fordonet och där utvärdera data i farten. Vi har beskrivit ett användningsfall men principen kan användas för godtyckligt användningsfall och det data som behövs för det.

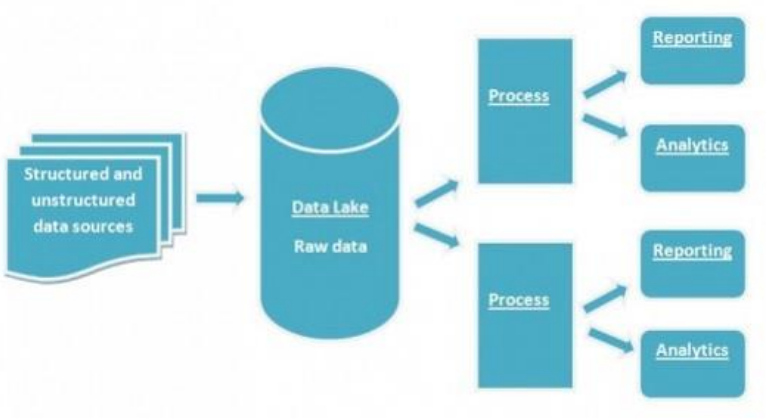
Arbetspaketets bidrag till FFI:s mål
I arbetspaketet har det tagits fram principer för kravställning, datadefinitioner och arkitektur för datainsamling. Resultaten är därmed direkt kopplade till *Fordons- och mobilitetstjänster*.

Arbetspaket 6 – Data warehouse

Arbetspaketet arbetade med hur externa användare ska få access och tillgång till ett företags data. Som användningsfall användes en extern användare inom projektet och arbetet resulterade i processer och guidelines som kan användas för liknande situationer i framtiden. I denna process följdes strikt GDPR och Scania ISEC-riktlinjer (Information Risk Assessment and Management (IRAM), Business Impact Assessment, Security Risk Assessment, etc.) innan data delades med externa användare. Data Privacy Officer har hjälpt till att utvärdera databehandlingen på grundval av gällande dataskyddslagstiftning och rekommenderat vissa riskreducerande åtgärder för att säkerställa att inte dela någon direkt eller indirekt personlig data. Från det delade datasetet uteslöts en del data och känslig data anonymiserades.

En stor del av arbetet var kopplat till Scanias Data Lake. Scania etablerade Data Lake för att samla data, till exempel information från uppkopplade fordon, försäljning, produktion och underhåll. En datasjö är en central lagringsplats som innehåller stora datamängder från många olika källor i ett rått, granulärt format. Den kan lagra strukturerad, semi-strukturerad eller

ostrukturerad data. Korrekt hantering av access och datatillgång för externa användare blir särskilt viktig på grund av att Scanias Data Lake innehåller data från så många olika källor.



Arbetspaketets bidrag till FFI:s mål
I arbetspaketet har det tagits fram processer och guidelines för att kunna dela data med externa parter samtidigt som GDPR och Scania ISEC-riktlinjer uppfylls. Därmed finns det en direkt koppling till *Regelverk, standardisering och juridik.*

7 Spridning och publicering

7.1 Kunskaps- och resultatspridning

Hur har/planeras projektresultatet att användas och spridas?	Markera med X	Kommentar
Öka kunskapen inom området	X	Resultaten har under projekttiden spridits inom projektgruppen genom kontinuerliga arbetsmöten och externt genom en stor mängd akademiska publikationer och ett antal genomförda workshops.
Föras vidare till andra avancerade tekniska utvecklingsprojekt	X	Lärdomar från projektet har till viss del nyttjats inom forskningsprojektet FAMOUS.
Föras vidare till produktutvecklingsprojekt	X	Resultaten används inom flera utvecklingsprojekt på Scania.
Introduceras på marknaden	X	Scania kommer introducera tjänster som till viss del baseras på resultat från projektet.
Användas i utredningar/regelverk/ tillståndsärenden/ politiska beslut		

7.2 Publikationer

Projektets publikationer listats under respektive arbetspaket nedan. Under arbetspaket 2 har även publikationer publicerade inom det kopplade föregående projektet IRIS tagits med.

Arbetspaket 2 – Datadrift och dynamisk data
Vetenskapliga artiklar

[1] S. Voronov, E. Frisk and M. Krysander, "*Data-Driven Battery Lifetime Prediction and Confidence Estimation for Heavy-Duty Trucks*" in IEEE Transactions on Reliability, vol. 67, no. 2, pp. 623-639, June 2018, DOI: https://doi.org/10.1109/TR.2018.2803798

[2] S. Voronov, D. Jung, E. Frisk, "*A forest-based algorithm for selecting informative variables using Variable Depth Distribution*", Engineering Applications of Artificial Intelligence, vol. 97, 2021, DOI: https://doi.org/10.1016/j.engappai.2020.104073
[3] S. Voronov, E. Frisk, and M. Krysander, "*Predictive Maintenance of Lead-Acid Batteries with Sparse Vehicle Operational Data*", International Journal of Prognostics and Health Management, vol 11, no. 1, 2020, DOI: https://doi.org/10.36001/ijphm.2020.v11i1.2608
[4] S. Voronov, E. Frisk, and M. Krysander, "*Lead-acid battery maintenance using multilayer perceptron models*", IEEE International Conference on Prognostics and Health Management. Seattle, USA, 2018, DOI: https://doi.org/10.1109/ICPHM.2018.8448472
[5] S. Voronov, D. Jung, E. Frisk, "*Variable selection for heavy-duty vehicle battery failure prognostics using random survival forests"*, *Third European Conference of the Prognostics and Health Management Society 2016*. Bilbao, Spain.
[6] Sergii Voronov, Daniel Jung, and Erik Frisk, "*Heavy-duty truck battery failure prognostics using random survival forests*", IFAC Advances in Automotive Control, 2016. Kolmården, Sweden. DOI: https://doi.org/10.1016/j.ifacol.2016.08.082

Avhandlingar
[7] S. Voronov, "*Data-driven lead-acid battery lifetime prognostics*", Licentiate thesis, Linköping University, 2017. Link: http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-137526
[8] S. Voronov, "*Machine Learning Models for Predictive Maintenance*", Doctoral thesis, Linköping University, 2020, Link: http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-162649

Examensarbeten
- E. Bremer, *"Prediction of Component Breakdowns in Commercial Trucks : Using Machine Learning on Operational and Repair History Data"*, Master Thesis, EECS, KTH, 2020
- F. Liljefors, *"Time dependent modeling of turbocharger failure using machine learning"*, Master Thesis, EECS, KTH, 2020

Arbetspaket 3 – Tolkningsbarhet av modeller

Vetenskapliga artiklar
[9] R. Gurung, T. Lindgren, and H. Boström, *"Learning Random Forest from Histogram Data Using Split Specific Axis Rotation"*, International Journal of Machine Learning and Computing vol. 8, no. 1, pp. 74-79, 2018
[10] T. Lindgren, *"Random Rule Sets – Combining Random Covering with the Random Subspace Method"*, International Journal of Machine Learning and Computing vol. 8, no. 1, pp. 8-13, 2018
[11] T. Lindgren, *"On Data Driven Organizations and the Necessity of Interpretable Models"*, Smart Grid and Internet of Things, Second EAI International Conference, (SGIoT), pp. 121-130, 2018
[12] H. Boström, R. Gurung, T. Lindgren, and U. Johansson, *"Explaining Random Forest Predictions with Association Rules"*, Archives of Data Science, Series A, vol. 5, no. 1, pp. 121-130, 2018
[13] T. Lindgren, P. Papapetrou, I. Samsten and L. Asker, *"Example-Based Feature Tweaking Using Random Forests"*, IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI), pp. 53-60, 2019
[14] R. Gurung, T. Lindgren, H. Boström, *"An Interactive Visual Tool Enhance Understanding of Random Forest Prediction"*, Multidisciplinary Facets of Data Science: Proceedings of the European Conference on Data Analysis 2019 in Archives of Data Science, Series A, (submitted)
[15] R. Gurung, *"NOx Sensor Failure Analysis in Heavy Trucks using Adapted Random Survival Forest for Histograms"*, In Proceedings of the 5th International Conference on Machine Learning, Optimization, and Data Science, 2019

[16] H. Boström, U. Johansson and A. Vesterberg, *"Predicting with Confidence from Survival Data."* Conformal and Probabilistic Prediction and Applications. PMLR, 2019
[17] H. Boström and U. Johansson, *"Mondrian conformal regressors."* Conformal and Probabilistic Prediction and Applications. PMLR, 2020
[18] L. Zed, A. Nicholas, P. Papapetrou, and T. Lindgren, *"Z-Hist: A Temporal Abstraction of Multivariate Histogram Snapshots"*, In Proceedings of the 19th Symposium on Intelligent Data Analysis, 2021

Avhandlingar

[19] R. Gurung, *"Random Forest for Histogram Data – An application in data-driven prognostic models for heavy-duty trucks"*, Doctoral Thesis, DSV, Stockholm University, Sweden, 2019

Examensarbeten

- B. Besimi, *"Evaluation of Methods for Handling Different Types of Concept Drifts"*, Master Thesis, DSV, Stockholm University, Sweden, 2018
- A. Schubert, *"Autoencoders for compressing histogram data"*, Bachelor Thesis, DSV, Stockholm University, Sweden, 2019
- A. Aparna and M. N. Chavez De La Vega, *"Techniques to Analyze Histogram Data"*, Master Thesis, DSV, Stockholm University, Sweden, 2020
- J. Bärlund, *"Handling missing feature values in histogram data"*, Master Thesis, DSV, Stockholm University, Sweden, 2020

Arbetspaket 4 – Driftsättning av prognostiska modeller

Examensarbeten

- J. van Miltenburg*, "Conformal survival predictions at a user-controlled time point: The introduction of time point specialized Conformal Random Survival Forests"*, Master Thesis, EECS, KTH, 2018
- I. Aparicio Vázquez, *"Venn Prediction for Survival Analysis: Experimenting with Survival Data and Venn Predictors"*, Master Thesis, EECS, KTH, 2020

8 Slutsatser och fortsatt forskning

Alla parter i projektet är nöjda med resultaten, samarbetet och kunskapsöverföringen. Forskningsfrågorna baserades på verkliga behov och data, och projektet lyckades kombinera vetenskaplig forskning med industriella implementeringar på en djupgående nivå.

De väsentliga projektmålen beskrivna i kapitel 5 har uppfyllts och därigenom har projektet avsevärt stärkt forskningsområdet inom maskininlärning och prognostisk – och då särskilt inom konceptdrift och tolkningsbarhet hos modeller. Projektet levererade två doktorsavhandlingar, runt 15 vetenskapliga artiklar, och åtta examensarbeten. Dessutom tog projektet fram programvara för tolkningsbarhet, ramverk/arkitektur för driftsättning av maskininlärningsmodeller, en demonstrator för datainsamling och processer/guidelines för datadelning med externa användare. Totalt hölls 17 workshops, varav ett var publikt, för kunskapsöverföring och spridning av resultaten.

En utmaning för de akademiska parterna var att det var svårare att få vetenskapliga artiklar accepterade eftersom andra forskare utanför projektet inte får tillgång till den data som används, och kan då inte verifiera resultaten i artiklarna. En projekttid på drygt tre år kan normalt sett vara lite kort eftersom många forskningsprojekt har ganska lång startsträcka, men i det här fallet funkade det bra eftersom CODA byggde vidare på forskningsprojektet IRIS.

Forskningsfrågor som uppkommit under projektets gång, men inte har kunnat behandlats fullständigt är exempelvis:

- Hantering av osäkerhet – hur avgör man om en maskininlärningsmodell är tillräckligt bra för produktionssättning?
- Hur blir ett företag datadrivet? Detta är en komplex fråga som förmodligen är starkt kopplad till ett företags digitala mognadsgrad.
- Hur kan fysiska och datadrivna modeller användas tillsammans? Data kommer inte bara från produkter i drift utan kan också genereras från fysiska modeller som används under produktutvecklingsfasen.

Alla dessa forskningsfrågor skulle kunna vara lämpliga för fortsatt forskning.

9 Deltagande parter och kontaktpersoner

Kontakta följande personer vid frågor gällande projektet:

- Projektledare: Jesper Brauer, Scania CV, jesper.brauer@scania.com
- Arbetspaket 2 – Datadrift och dynamisk data, Erik Frisk, LiU, erik.frisk@liu.se
- Arbetspaket 3 – Tolkningsbarhet av modeller, Tony Lindgren, SU, tony@dsv.su.se
- Arbetspaket 4 – Driftsättning av prognostiska modeller, Anders Vesterberg, Scania CV, anders.vesterberg@scania.com
- Arbetspaket 5 – Effektiviserad datainsamling, Pär Sundbäck, Scania CV, par.sundback@scania.com
- Arbetspaket 6 – Data warehouse, Henrik Brandin, Scania CV, henrik.brandin@scania.com